Color Science

Robert M. Boynton
University of California, San Diego

GLOSSARY

Chromaticity Ratios x, y, z of each of the tristimulus values of a light to the sum of the three tristimulus values X, Y, Z, these being the amounts of three primaries required to match the color of the light.

Chromaticity diagram Plane diagram formed by plotting one of the three chromaticity coordinates against another (usually y versus x).

Color Characteristics of sensations elicited by light by which a human observer can distinguish between two structure-free patches of light of the same size and shape.

Colorant A substance, such as a dye or pigment, that modifies the color of objects or imparts color to otherwise achromatic objects.

Colorimetry Measurement and specification of color.

Color matching Action of making a test color appear the same as a reference color.

Color order System of reference whereby the relation of one color to another can be perceived and the position of that color can be established with respect to the universe of all colors.

Color rendering General expression for the effect of a light source on the color appearance of objects in comparison with their color appearance under a reference light source.

Color temperature Absolute temperature of a blackbody radiator having a chromaticity closest to that of a light source being specified.

Metamerism (1) Phenomenon whereby lights of different spectral power distributions appear to have the same color. (2) Degree to which a material appears to change color when viewed under different illuminants.

Optimal colors Stimuli that for a given chromaticity have the greatest luminous reflectance.

Primaries (1) Additive: Any one of three lights in terms of which a color is specified by giving the amount of each required to match it by combining the lights. (2) Subtractive: Set of dyes or pigments that, when mixed in various proportions, provides a gamut of colors.

Radiance Radiant flux per unit solid angle (intensity) per unit area of an element of an extended source or reflecting surface in a specified direction.

Reflectance Ratio of reflected to incident light.

Reflection Process by which incident flux leaves a surface or medium from the incident side, without change in wavelength.

COLOR SCIENCE examines a fundamental aspect of human perception. It is based on experimental study under controlled conditions susceptible to physical measurement. For a difference in color to be perceived between two surfaces, three conditions must be satisfied: (1) There must be an appropriate source of illumination, (2) the two surfaces must not have identical spectral reflectances, and (3) an observer must be present to view them. This article is concerned with the relevant characteristics of lights, surfaces, and human vision that conjoin to allow the perception of object color.

I. PHYSICAL BASIS OF PERCEIVED COLOR

The physical basis of color exists in the interaction of light with matter, both outside and inside the eye. The sensation of color depends on physiological activity in the visual system that begins with the absorption of light in photoreceptors located in the retina of the eye and ends with patterns of biochemical activity in the brain. Perceived color can be described by the color names white, gray, black, yellow, orange, brown, red, green, blue, purple, and pink. These 11 basic color terms have unambiguous referents in all fully developed languages. All of these names (as well as combinations of these and many other less precisely used nonbasic color terms) describe colors, but white, gray, and black are excluded from the list of those called hues. Colors with hue are called chromatic colors; those without are called achromatic colors.

Although color terms are frequently used in reference to all three aspects of color (e.g., one may speak of a sensation of red, a red surface, or a red light), such usage is scientifically appropriate only when applied to the sensation; descriptions of lights and surfaces should be provided in physical and geometrical language.

II. CIE SYSTEM OF COLOR SPECIFICATION

A. Basic Color-Matching Experiment

The most fundamental experiment in color science entails the determination of whether two fields of light such as those that might be produced on a screen with two slide projectors, appear the same or different. If such fields are abutted and the division between them disappears to form a single, homogeneous field, the fields are said to match. A match will, of course, occur if there is no physical difference between the fields, and in special cases color matches are also possible when substantial physical differences exist between the fields. An understanding of how this can happen provides an opening to a scientific understanding of this subject.

Given an initial physical match, a difference in color can be introduced by either of two procedures, which are often carried out in combination. In the first instance, the radiance of one part of a homogeneous field is altered without any change in its relative spectral distribution. This produces an achromatic color difference. In the second case, the relative spectral distribution of one field is changed such that, for all possible relative radiances of the two fields, no match is possible. This is called a chromatic color difference.

When fields of different spectral distributions can be adjusted in relative radiance to eliminate all color difference, the result is termed a metameric color match. In a color-matching experiment, a test field is presented next to a comparison field and the observer causes the two fields to match exactly by manipulating the radiances of so-called primaries provided to the comparison field. Such primaries are said to be added; this can be accomplished by superposition with a half-silvered mirror, by superimposed images projected onto a screen, by very rapid temporal alternation of fields at a rate above the fusion frequency for vision, or by the use of pixels too small and closely packed to be discriminated (as in color television). If the primaries are suitably chosen (no one of them should be matched by any possible mixture of the other two), a human observer with normal color vision can uniquely match any test color by adjusting the radiances of three monochromatic primaries. To accomplish this, it sometimes proves necessary to shift one of the primaries so that it is added to the color being matched; it is useful to treat this as a negative radiance of that primary in the test field. The choice of exactly three primaries is by no means arbitrary: If only one or two primaries are used, matches are generally impossible, whereas if four or more primaries are allowed, matches are not uniquely determined.

The result of the color-matching experiment can be represented mathematically as $t(T) = r(R) + g(G) + b(B)$, meaning that t units of test field T produce a color that is matched by an additive combination of r units of primary R, g units of primary G, and b units of primary B, where one or two of the quantities r, g, or b may be negative. Thus any color can be represented as a vector in R, G, B space. For small, centrally fixated fields, experiment shows that the transitive, reflexive, linear, and associative properties of algebra apply also to their empirical counterparts, so that color-matching equations can be manipulated to predict matches that would be made with a

change in the choice of primaries. These simple relations break down for very low levels of illumination and also with higher levels if the fields are large enough to permit significant contributions by rod photoreceptors or if the fields are so bright as to bleach a significant fraction of cone photopigments, thus altering their action spectra.

Matches are usually made by a method of adjustment, an iterative, trial-and-error procedure whereby the observer manipulates three controls, each of which monotonically varies the radiance of one primary. Although such settings at the match point may be somewhat more variable than most purely physical measurements, reliable data result from the means of several settings for each condition tested. A more serious problem, which will not be treated in this article, results from differences among observers. Although not great among those with normal color vision, such differences are by no means negligible. (For those with abnormal color vision, they can be very large.) To achieve a useful standardization—one that is unlikely to apply exactly to any particular individual—averages of normal observers are used, leading to the concept of a standard observer.

In the color-matching experiment, an observer is in effect acting as an analog computer, solving three simultaneous equations by iteration, using his or her sensations as a guide. Although activity in the brain underlies the experience of color, the initial encoding of information related to wavelength is in terms of the ratios of excitations of three different classes of cone photoreceptors in the retina of the eye, whose spectral sensitivities overlap. Any two physical fields, whether of the same or different spectral composition, whose images on the retina excite each of the three classes of cones in the same way will be indiscriminable. The action spectra of the three classes of cones in the normal eye are such that no two wavelengths in the spectrum produce exactly the same ratios of excitations among them.

B. Imaginary Primaries

Depending on the choice of primaries, many different sets of color-matching functions are possible, all of which describe the same color-matching behavior. Figure 1 shows experimental data for the primaries 435.8, 546.1, and

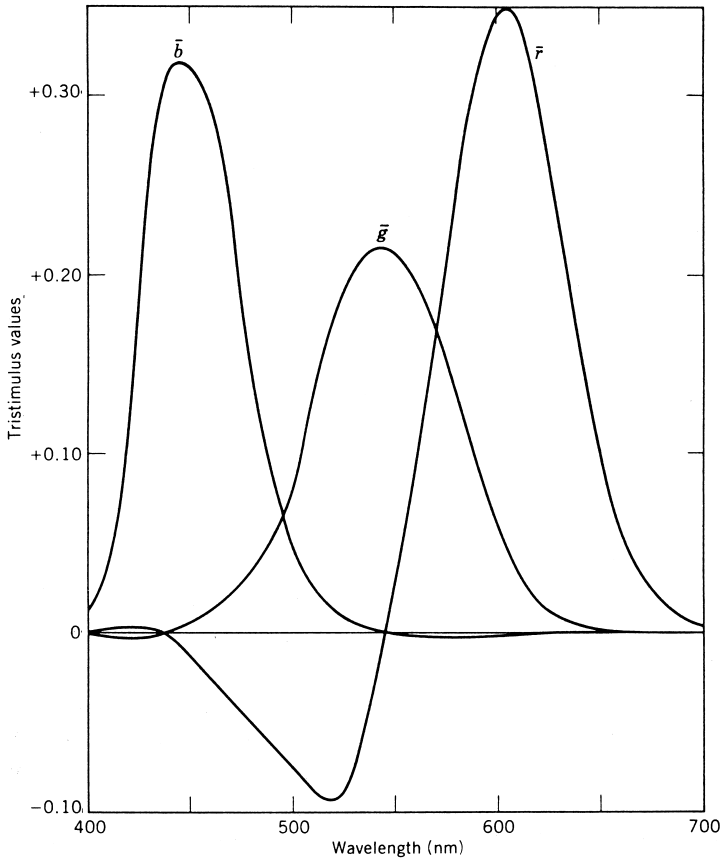


FIGURE 1 Experimental color-matching data for primaries at 435.8, 546.1, and 700.0 nm. [From Billmeyer, F. W., Jr., and Saltzmann, M. (1981). "Principles of Color Technology," 2nd ed. Copyright ©1981 John Wiley & Sons, Inc. Reprinted by permission of John Wiley & Sons, Inc.]

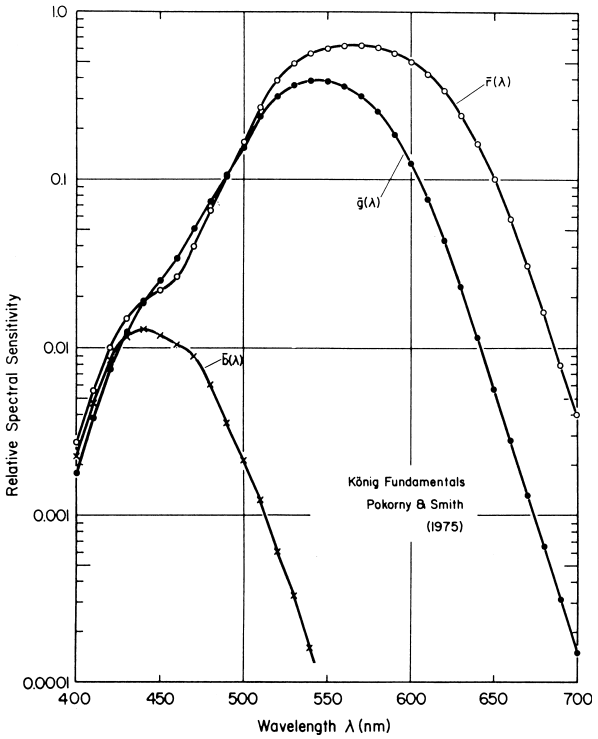


FIGURE 2 Estimates of human cone action spectra (König fundamentals) derived by V. Smith and J. Pokorny. [From Wyszecki, G., and Stiles, W. S. (1982). "Color Science: Concepts and Methods, Quantitative Data and Formulate," 2nd ed. Copyright ©1982 John Wiley & Sons, Inc. Reprinted by permission of John Wiley & Sons, Inc.]

700.0 nm. Depicted in Fig. 2 are current estimates of the spectral sensitivities of the three types of cone photoreceptors. These functions, which have been inferred from the data of psychophysical experiments of various kinds, agree reasonably well with direct microspectrophotometric measurements of the absorption spectra of outer segments of human cone photoreceptors containing the photopigments that are the principal determinants of the spectral sensitivity of the cones.

The cone spectral sensitivities may be regarded as color-matching functions based on primaries that are said to be imaginary in the sense that, although calculations of color matches based on them are possible, they are not physically realizable. To exist physically, each such primary would uniquely excite only one type of cone, whereas real primaries always excite at least two types.

Another set of all-positive color-matching functions, based on a different set of imaginary primaries, is given in Fig. 3. This set, which makes very similar predictions about color matches as the cone sensitivity curves, was adopted as a standard by the International Commission on Illumination (CIE) in 1931.

By simulating any of these sets of sensitivity functions in three optically filtered photocells, it is possible to remove the human observer from the system of color measurement (colorimetry) and develop a purely physical (though necessarily very limited) description of color, one that can be implemented in automated colorimeters.

C. Chromaticity Diagram

A useful separation between the achromatic and chromatic aspects of color was achieved in a system of colorimetry adopted by the CIE in 1931. This was the first specification of color to achieve international agreement; it remains today the principal system used internationally for specifying colors quantitatively, without reference to a set of actual samples.

The color-matching functions $\bar{x}(\lambda)$, $\bar{y}(\lambda)$, and $\bar{z}(\lambda)$ are based on primaries selected and smoothed to force the $\bar{y}(\lambda)$ function to be proportional to the spectral luminous efficiency function $V(\lambda)$, which had been standardized a decade earlier to define the quantity of "luminous flux" in lumens per watt of radiant power. The $\bar{x}(\lambda)$, $\bar{y}(\lambda)$, and $\bar{z}(\lambda)$ functions were then scaled to equate the areas under the curves, an operation that does not alter the predictions they make about color matches.

To specify the color of a patch of light, one begins by integrating its spectral radiance distribution $S(\lambda)$ in turn with the three color-matching functions:

$$X = k \int S(\lambda)\bar{x}(\lambda)\, d\lambda,$$

$$Y = k \int S(\lambda)\bar{y}(\lambda)\, d\lambda,$$

$$Z = k \int S(\lambda)\bar{z}(\lambda)\, d\lambda.$$

The values X, Y, and Z are called relative tristimulus values; these are equal for any light having an equal-radiance spectrum. Tristimulus values permit the specification of color in terms of three variables that are related to cone sensitivities rather than by continuous spectral radiance distributions, which do not. Like R, G, and B, the tristimulus values represent the coordinates of a three-dimensional vector whose angle specifies chromatic color and whose length characterizes the amount of that color.

Chromaticity coordinates, which do not depend on the amount of a color, specify each of the tristimulus values relative to their sum:

$$x = X/(X + Y + Z);$$

$$y = Y(X + Y + Z);$$

$$z = Z/(X + Y + Z)$$

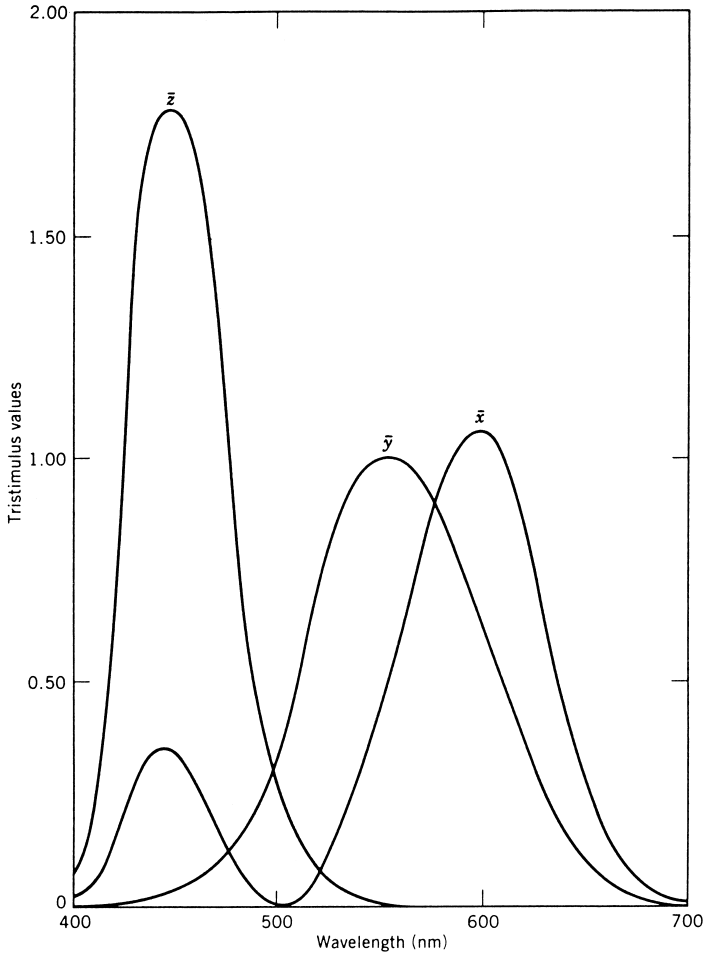


FIGURE 3 Tristimulus values of the equal-energy spectrum of the 1931 CIE system of colorimetry. [From Billmeyer, F. W., Jr., and Saltzmann, M. (1981). "Principles of Color Technology," 2nd ed. Copyright ©1981 John Wiley & Sons, Inc. Reprinted by permission of John Wiley & Sons, Inc.]

Given any two of these, the third is determined (e.g., $z = 1 - x - y$). Therefore, full information about chromaticity can be conveniently represented in a two-dimensional diagram, with y versus x having been chosen by the CIE for this purpose. The resulting chromaticity diagram is shown in Fig. 4. If one wishes to specify the quantity of light as well, the Y tristimulus value can be given, allowing a color to be fully specified as x, y, and Y, instead of X, Y, and Z. The manner in which the quantity of light Y is specified is determined by the normalization constant k.

Depending on the choice of primaries for determining color-matching functions, many other chromaticity diagrams are possible. For example, the set of color-matching functions of Fig. 1 leads to the chromaticity diagram of Fig. 5. This so-called *RGB* system is seldom used.

The affine geometry of chromaticity diagrams endows all of them with a number of useful properties. Most fundamental is that an additive mixture of any two lights will fall along a straight line connecting the chromaticities of the mixture components. Another is that straight lines on one such diagram translate into straight lines on any other related to it by a change of assumed primaries. The locations of the imaginary primaries X, Y, and Z are shown in Fig. 5, where one sees that the triangle formed by them completely encloses the domain of realizable colors. The lines X–Y and X–Z of Fig. 5 form the coordinate axes of the CIE chromaticity diagram of Fig. 4. Conversely, the lines B–G and B–R in Fig. 4 form the coordinate axes of the chromaticity diagram of Fig. 5. The uneven grid of nonoorthogonal lines in Fig. 4, forming various angles at their intersections, translates into the regular grid of evenly spaced, orthogonal lines in Fig. 5. This illustrates that angles and areas have no instrinsic meaning in chromaticity diagrams.

The CIE in 1964 adopted an alternative set of color-matching functions based on experiments with large (10°) fields. Their use is recommended for making predictions about color matches for fields subtending more than 4° at the eye.

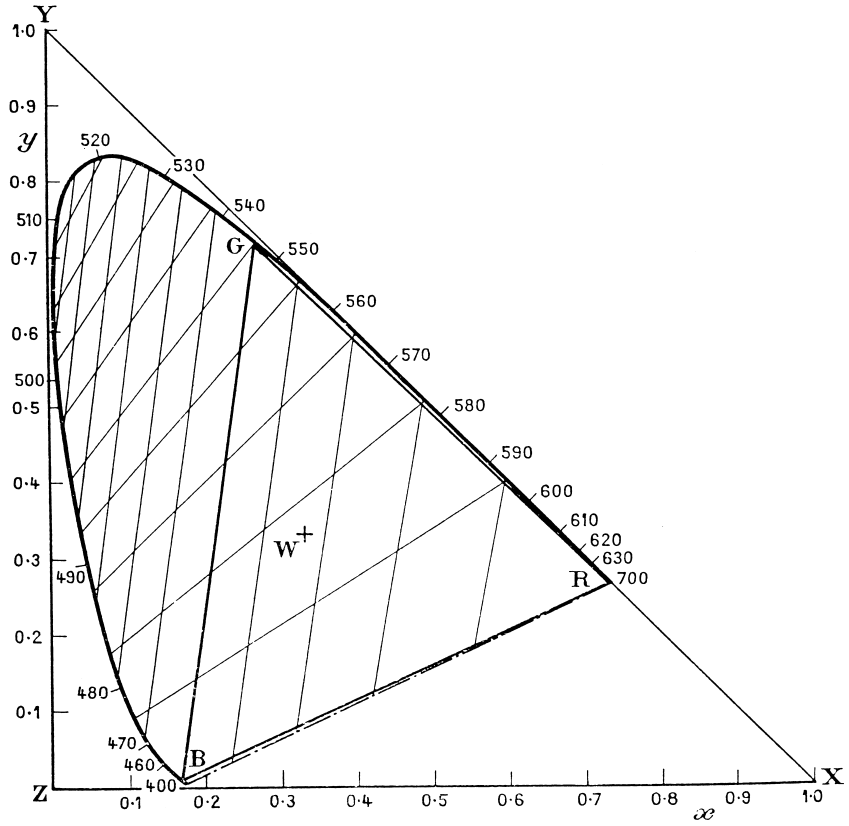


FIGURE 4 The *XYZ* chromaticity diagram showing the locations of the *RGB* primaries of Fig. 5 and the projection of the rectilinear grid of that figure onto this one. [From LeGrand, Y. (1957). "Light, Colour, and Vision," 2nd ed., Wiley Interscience, New York.]

Table I lists values of the CIE color-matching functions for 2° and 10° fields at 10-nm wavelength values. Tables for 1-nm wavelength values for 2° and 10° fields are available in *Color Measurement*, the second volume in the series *Optical Radiation Measurements*, edited by F. Grum and C. J. Bartleson.

III. COLOR RENDERING

From an evolutionary viewpoint, it is not surprising that sunlight is an excellent source for color rendering. Its strong, gap-free spectral irradiance distribution (Fig. 6) allows the discrimination of a very large number of surface-color differences. Color appearance in sunlight provides the standard against which the adequacy of other sources for color rendering is often judged.

A. Best and Worst Artificial Sources for Color Rendering

Of the light sources in common use today, low-pressure sodium is one of the poorest for color rendering, coming very close to being one of the worst possible. This illuminant consists mainly of the paired sodium lines that lie very close together (at 589.0 and 589.6 nm) in the "yellow" region of the spectrum; although some other spectral lines are also represented, these are present at such low relative radiances that low-pressure sodium lighting is for practical purposes monochromatic.

For a surface that does not fluoresce, its spectral reflectance characteristics can modify the quantity and geometry of incident monochromatic light, but not its wavelength. Viewed under separate monochromatic light sources of the same wavelength, any two surfaces with arbitrarily chosen spectral distributions can be made to match, both physically and visually, by adjusting the relative radiances of incident lights. Therefore, no chromatic color differences can exist under monochromatic illumination.

The best sources for color rendering emit continuous spectra throughout the visible region. Blackbody radiation, which meets this criterion, is shown for three temperatures in Fig. 7. These curves approximate those for tungsten sources at these temperatures.

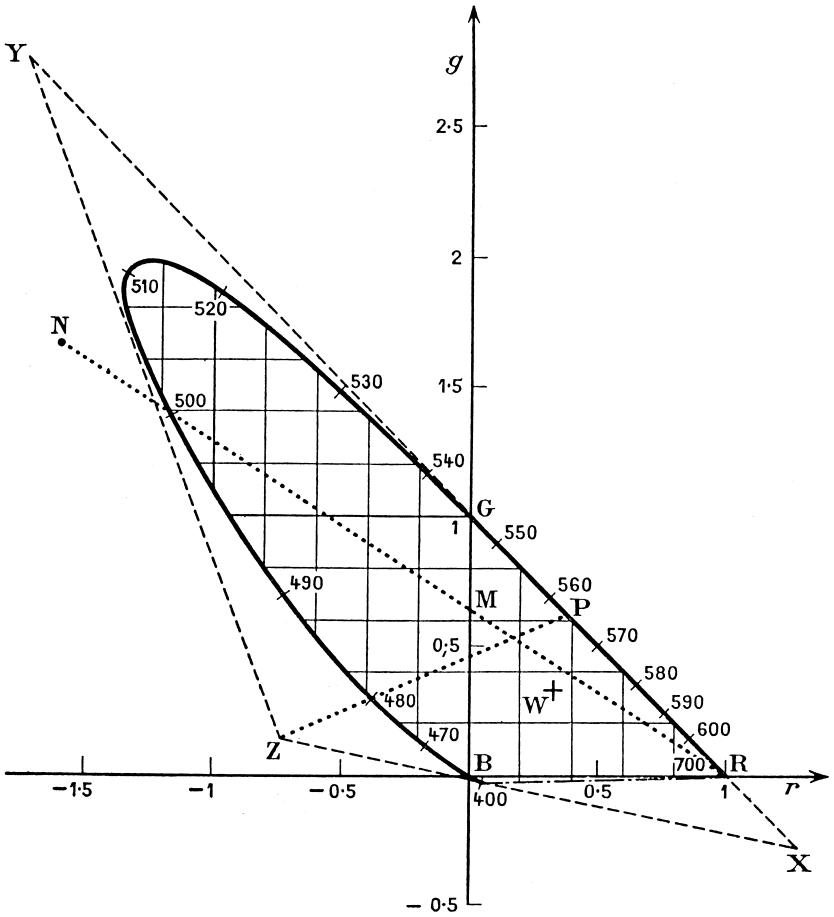


FIGURE 5 The *RGB* chromaticity diagram showing the locations of the *XYZ* primaries of Fig. 4. [From LeGrand, Y. (1957). "Light, Colour, and Vision," 2nd ed., Wiley Interscience, New York.]

B. Intermediate Quality of Fluorescent Lighting

Much of the radiant flux produced by incandescence emerges as infrared radiation at wavelengths longer than those visible; this is not true of fluorescent light, which is more efficiently produced, accounting for its widespread use. This light results from the electrical energizing of mercury vapor, which emits ultraviolet radiation. Although itself invisible, this radiation elicits visible light by causing the fluorescence of phosphors suspended in a layer coating the inside of a transparent tube.

Fluorescent lamps emit energy at all visible wavelengths, which is a good feature for color rendering, but their spectra are punctuated by regions of much higher radiance whose spectral locations depend on the phosphors chosen and the visible radiations of mercury vapor. Radiant power distributions of six types of fluorescent lamps are shown in Fig. 8.

C. Efficacy

The amount of visible light emitted by a source is measured in lumens, determined by integrating its radiant power output $S(\lambda)$ with the spectral luminous efficiency function $V(\lambda)$. The latter, which is proportional to $\bar{y}(\lambda)$, peaks at ~555 nm. Therefore, the theoretically most efficient light source would be monochromatic at this wavelength, with the associated inability to render chromatic color differences. Efficacy does not include power lost in the conversion from electrical input to radiant output, which may vary independently of the efficacy of the light finally produced.

D. Correlated Color Temperature

A blackbody, or Planckian radiator, is a cavity within a heated material from which heat cannot escape. No matter what the material, the walls of the cavity exhibit a

TABLE I Spectral Tristimulus Values for Equal Spectral Power Source

a. CIE 1931 Standard Observer				b. CIE 1964 Supplementary Observer			
Wavelength (nanometer)	$\bar{x}(\lambda)$	$\bar{y}(\lambda)$	$\bar{z}(\lambda)$	Wavelength (nanometer)	$\bar{x}_{10}(\lambda)$	$\bar{y}_{10}(\lambda)$	$\bar{z}_{10}(\lambda)$
380	0.0014	0.0000	0.0065	380	0.0002	0.0000	0.0007
385	0.0022	0.0001	0.0105	385	0.0007	0.0001	0.0029
390	0.0042	0.0001	0.0201	390	0.0024	0.0003	0.0105
395	0.0076	0.0002	0.0362	395	0.0072	0.0008	0.0323
400	0.0143	0.0004	0.0679	400	0.0191	0.0020	0.0860
405	0.0232	0.0006	0.1102	405	0.0434	0.0045	0.1971
410	0.0435	0.0012	0.2074	410	0.0847	0.0088	0.3894
415	0.0776	0.0022	0.3713	415	0.1406	0.0145	0.6568
420	0.1344	0.0040	0.6456	420	0.2045	0.0214	0.9725
425	0.2148	0.0073	1.0391	425	0.2647	0.0295	1.2825
430	0.2839	0.0116	1.3856	430	0.3147	0.0387	1.5535
435	0.3285	0.0618	1.6230	435	0.3577	0.0496	1.7985
440	0.3483	0.0230	1.7471	440	0.3837	0.0621	1.9673
445	0.3481	0.0298	1.7826	445	0.3687	0.0747	2.0273
450	0.3362	0.0380	1.7721	450	0.3707	0.0895	1.9948
455	0.3187	0.0480	1.7441	455	0.3430	0.1063	1.9007
460	0.2908	0.0600	1.6692	460	0.3023	0.1282	1.7454
465	0.2511	0.0739	1.5281	465	0.2541	0.1528	1.5549
470	0.1954	0.0910	1.2876	470	0.1956	0.1852	1.3176
475	0.1421	0.1126	1.0419	475	0.1323	0.2199	1.0302
480	0.0956	0.1390	0.8130	480	0.0805	0.2536	0.7721
485	0.0580	0.1693	0.6162	485	0.0411	0.2977	0.5701
490	0.0320	0.2080	0.4652	490	0.0162	0.3391	0.4153
495	0.0147	0.2586	0.3533	495	0.0051	0.3954	0.3024
500	0.0049	0.3230	0.2720	500	0.0038	0.4608	0.2185
505	0.0024	0.4073	0.2123	505	0.0154	0.5314	0.1592
510	0.0093	0.5030	0.1582	510	0.0375	0.6067	0.1120
515	0.0291	0.6082	0.1117	515	0.0714	0.6857	0.0822
520	0.0633	0.7100	0.0782	520	0.1177	0.7618	0.0607
525	0.1096	0.7932	0.0573	525	0.1730	0.8233	0.0431
530	0.1655	0.8620	0.0422	530	0.2365	0.8752	0.0305
535	0.2257	0.9149	0.0298	535	0.3042	0.9238	0.0206
540	0.2904	0.9540	0.0203	540	0.3768	0.9620	0.0137
545	0.3597	0.9803	0.0134	545	0.4516	0.9822	0.0079
550	0.4334	0.9950	0.0087	550	0.5298	0.9918	0.0040
555	0.5121	1.0000	0.0057	555	0.6161	0.9991	0.0011
560	0.5945	0.9950	0.0039	560	0.7052	0.9973	0.0000
565	0.6784	0.9786	0.0027	565	0.7938	0.9824	0.0000
570	0.7621	0.9520	0.0021	570	0.8787	0.9556	0.0000
575	0.8425	0.9154	0.0018	575	0.9512	0.9152	0.0000
580	0.9163	0.8700	0.0017	580	1.0142	0.8689	0.0000
585	0.9786	0.8163	0.0014	585	1.0743	0.8526	0.0000
590	1.0263	0.7570	0.0011	590	1.1185	0.7774	0.0000
595	1.0567	0.6949	0.0010	595	1.1343	0.7204	0.0000
600	1.0622	0.6310	0.0008	600	1.1240	0.6583	0.0000
605	1.0456	0.5668	0.0006	605	1.0891	0.5939	0.0000
610	1.0026	0.5030	0.0003	610	1.0305	0.5280	0.0000

continues

TABLE I (*continued*)

a. CIE 1931 Standard Observer				b. CIE 1964 Supplementary Observer			
Wavelength (nanometer)	$\bar{x}(\lambda)$	$\bar{y}(\lambda)$	$\bar{z}(\lambda)$	**Wavelength (nanometer)**	$\bar{x}_{10}(\lambda)$	$\bar{y}_{10}(\lambda)$	$\bar{z}_{10}(\lambda)$
615	0.9384	0.4412	0.0002	615	0.9507	0.4618	0.0000
620	0.8544	0.3810	0.0002	620	0.8563	0.3981	0.0000
625	0.7514	0.3210	0.0001	625	0.7549	0.3396	0.0000
630	0.6424	0.2650	0.0000	630	0.6475	0.2835	0.0000
635	0.5419	0.2170	0.0000	635	0.5351	0.2283	0.0000
640	0.4479	0.1750	0.0000	640	0.4316	0.1798	0.0000
645	0.3608	0.1382	0.0000	645	0.3437	0.1402	0.0000
650	0.2835	0.1070	0.0000	650	0.2683	0.1076	0.0000
655	0.2187	0.0816	0.0000	655	0.2043	0.0812	0.0000
660	0.1649	0.0610	0.0000	660	0.1526	0.0603	0.0000
665	0.1212	0.0446	0.0000	665	0.1122	0.0441	0.0000
670	0.0874	0.0320	0.0000	670	0.0813	0.0318	0.0000
675	0.0636	0.0232	0.0000	675	0.0579	0.0226	0.0000
680	0.0468	0.0170	0.0000	680	0.0409	0.0159	0.0000
685	0.0329	0.0119	0.0000	685	0.0286	0.0111	0.0000
690	0.0227	0.0082	0.0000	690	0.0199	0.0077	0.0000
695	0.0158	0.0057	0.0000	695	0.0318	0.0054	0.0000
700	0.0114	0.0041	0.0000	700	0.0096	0.0037	0.0000
705	0.0081	0.0029	0.0000	705	0.0066	0.0026	0.0000
710	0.0058	0.0021	0.0000	710	0.0046	0.0018	0.0000
715	0.0041	0.0015	0.0000	715	0.0031	0.0012	0.0000
720	0.0029	0.0010	0.0000	720	0.0022	0.0008	0.0000
725	0.0020	0.0007	0.0000	725	0.0015	0.0006	0.0000
730	0.0014	0.0005	0.0000	730	0.0010	0.0004	0.0000
735	0.0010	0.0004	0.0000	735	0.0007	0.0003	0.0000
740	0.0007	0.0002	0.0000	740	0.0005	0.0002	0.0000
745	0.0005	0.0002	0.0000	745	0.0004	0.0001	0.0000
750	0.0003	0.0001	0.0000	750	0.0003	0.0001	0.0000
755	0.0002	0.0001	0.0000	755	0.0001	0.0001	0.0000
760	0.0002	0.0001	0.0000	760	0.0001	0.0000	0.0000
765	0.0002	0.0001	0.0000	765	0.0001	0.0000	0.0000
770	0.0001	0.0000	0.0000	770	0.0001	0.0000	0.0000
775	0.0001	0.0000	0.0000	775	0.0000	0.0000	0.0000
780	0.0000	0.0000	0.0000	780	0.0000	0.0000	0.0000
Totals	21.3714	21.3711	21.3715	Totals	23.3294	23.3324	23.3343

characteristic spectral emission, which is a function of its temperature. The locus of the chromaticity coordinates corresponding to blackbody radiation, as a function of temperature, plots in the chromaticity diagram as a curved line known as the Planckian locus (see Fig. 4). The spectral distribution of light from sources with complex spectra does not approximate that of a Planckian radiator. Nevertheless, it is convenient to have a single index by which to characterize these other sources of artificial light. For this purpose the CIE has defined a correlated color temperature, determined by calculating the chromaticity coordinates of the source and then locating the point on the blackbody locus perceptually closest to these coordinates.

E. Color-Rendering Index

The CIE has developed a system for attempting to specify the quality of color rendering supplied by any light source. The calculations are based on a set of reflecting samples specified in terms of their reflectance functions. The

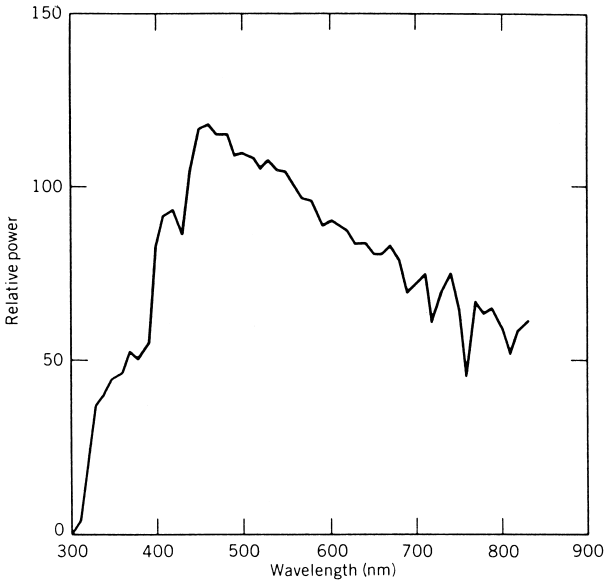


FIGURE 6 Spectral power distribution of typical daylight. [From Billmeyer, F. W., Jr., and Saltzmann, M. (1981). "Principles of Color Technology," 2nd ed., Copyright ©1981 John Wiley & Sons, Inc. Reprinted by permission of John Wiley & Sons, Inc.]

calculations begin with the choice of a reference illuminant specified as a blackbody (or daylight) radiator having a color temperature (or correlated color temperature) as close as possible to the correlated color temperature of the test illuminant; the choice of reference illuminant depends on the correlated color temperature of the test illuminant (daylight is used as a reference above 5000 K). For each of the samples, defined by their spectral reflectance functions, the amount of color shift ΔE introduced in going from reference to test illuminant is determined using the CIELUV formula described in Section VIB. There are 14 reference samples in all. A special color-rendering index R_i, peculiar to each sample, is calculated as $100–4.6\Delta E$. Most commonly a single-number index is calculated from the mean of a subset of eight special color-rendering indices to provide a final value known as the general color-rendering index R_a. The factor 4.6 was chosen so that a standard warm white fluorescent lamp would have an R_a of ~50; tungsten-incadescent sources score very close to 100. Table II gives R_a values for several commonly used artificial sources. Despite its official status, R_a is of limited value because of its many arbitrary features, especially its dependence on so limited a set of color samples. It is most useful for distinguishing large differences in color rendering, but not so useful for discriminating among sources of very high color-rendering properties. Individual values of R_i can be useful for determining the manner in which light sources differ in their color-rendering properties.

The intermediate color-rendering properties of most fluorescent light sources are closer to the best than to the worst. Mercury vapor and high-pressure sodium sources, widely used for street lighting, have poor color-rendering properties that fall between those of fluorescent and low-pressure sodium illumination.

IV. GLOBAL SURFACE PROPERTIES

The term *reflection* characterizes any of a variety of physical processes by which less than 100% of the radiant energy incident on a body at each wavelength is returned without change of wavelength. Reflection is too complicated for detailed specification at a molecular level for most surfaces and wavelengths of light. For this reason and because the molecular details are unimportant for many practical purposes, methods have been devised for measuring the spectral reflectance of a surface—the spectral distribution of returned light relative to that which is incident. Reflectance depends on the wavelength and angle of incidence of the light, as well as the angle(s) at which reflected light is measured.

A. Specular and Diffuse Reflectance

A familiar example of specular reflectance is provided by a plane mirror, in which the angles of light incidence and reflectance are equal. An ideal mirror reflects all incident light nonselectively with wavelength. If free of dust and suitably framed, the surface of an even less than ideal real mirror is not perceived at all; instead, the virtual image of an object located physically in front of the mirror is seen as if positioned behind.

Although specular reflectance seldom provides information about the color of a surface, there are exceptions. In particular, highly polished surfaces of metals such as gold, steel, silver, and copper reflect specularly. They also reflect diffusely from within but do so selectively with wavelength so that the specular reflection is seen to be tinged with the color of the diffuse component. More often, because highlights from most surfaces do not alter the spectral distribution of incident light, specular reflection provides information about the color of the source of light rather than that of the surface.

Diffuse reflectance, on the other hand, is typically selective with wavelength, and for the normal observer under typical conditions of illumination it is the principal determinant of the perceived color of a surface. A surface exhibiting perfectly diffuse reflectance returns all of the incident light with the distribution shown in Fig. 9, where the luminance (intensity per unit area) of the reflected light decreases a cosine function of the angle of reflection

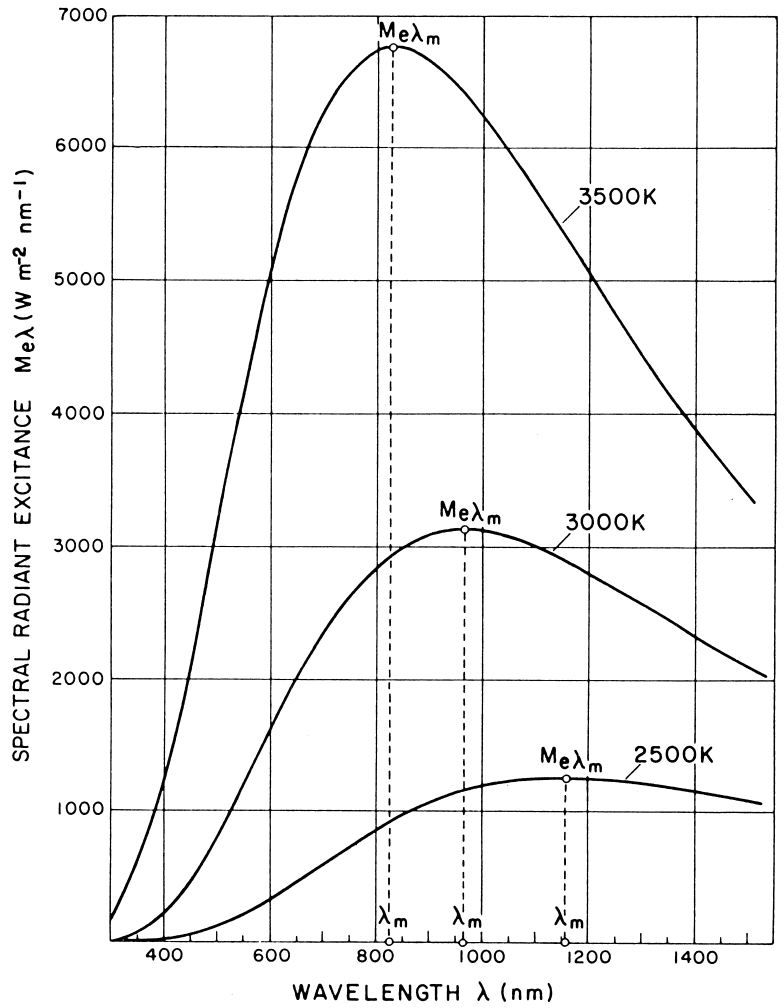


FIGURE 7 Spectral radiance distributions of a blackbody radiator at three temperatures. [From Wyszecki, G., and Stiles, W. S. (1982). "Color Science: Concepts and Methods, Quantitative Data and Formulae," 2nd ed. Copyright © 1982 John Wiley & Sons, Inc. Reprinted by permission of John Wiley & Sons, Inc.]

relative to normal. As such a surface is viewed more and more obliquely through an aperture, a progressively larger area of the surface fills the aperture—also a cosine function. The two effects cancel, causing the luminance of the surface and its subjective counterpart, lightness, to be independent of the angle of view.

No real surface behaves in exactly this way, although some surfaces approach it. Some simultaneously exhibit specular and diffuse reflectance; that of a new automobile provides a familiar example. The hard, highly polished outer surface exhibits specular reflectance of some of the incident light. The remainder is refracted into the layers below, which contain diffusely reflecting, spectrally selective absorptive pigments suspended in a binding matrix. Light not absorbed is scattered within this layer with an intensity pattern that may approximate that of a perfectly diffuse reflector. Because of the absorptive properties of the pigments, some wavelengths reflect more copiously than others, providing the physical basis for the perceived color of the object.

Many intermediate geometries are possible, which give rise to sensations of sheen and gloss; these usually enable one to predict the felt hardness or smoothness of surface without actually touching them.

B. Measuring Diffuse Surface Reflectance

The diffuse spectral reflectance of a surface depends on the exact conditions of measurement. To some extent these are arbitrary, so that in order for valid comparisons of measurements to be made among different laboratories and manufacturers, standard procedures are necessary. To agree on and specify such procedures has been one of the functions of the CIE, which has recommended four

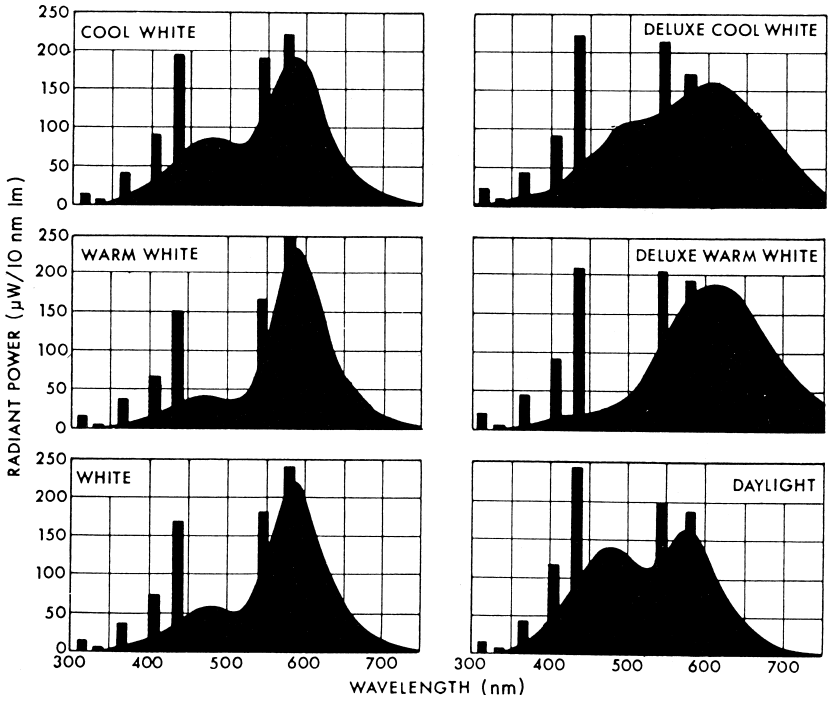


FIGURE 8 Spectral radiance distributions of six typical fluorescent lamps. [From Kaufman, J. E., ed. (1981). "IES Lighting Handbook; Reference Volume," © 1981 Illuminating Engineering Society of North America.]

procedures for measuring diffuse spectral reflectance, the most sophisticated of which is illustrated at the bottom left in Fig. 10. It makes use of an integrating sphere painted inside with a highly reflecting and spectrally nonselective paint made from barium sulfate. When light is admitted into an ideal integrating sphere, the sphere "lights up" uniformly as a result of multiple diffuse internal reflections. The size of the sphere does not matter so long as the ports cut into it, which permit entry and exit of the incident and reflected light, do not exceed 10% of the total surface area.

The surface to be measured fills an opening at the bottom, oriented horizontally. The incident light, which should be of limited cross section in order to be confined to the sample being measured, enters at an angle of 5° to normal. To eliminate the specular component of reflection from the measurement, a light trap is introduced, centered at an angle of 5° on the other side of normal; the remaining diffusely reflected component illuminates the sphere. Ideally, the exit port could be located almost anywhere. In practice, it is located as shown, so that it "sees" only a small opposite section of the sphere. As an added precaution, a small baffle is introduced to block the initially reflected component of light, which otherwise would strike the sphere in the area being measured. When the cap, shown in the lower left of Fig. 10 is black, it eliminates the specular (or direct) component, and only the diffuse reflectance is measured. When the cap is white (or when the sphere's surface is continuous, as at the bottom right of the figure), both specular and diffuse components contribute, and the measurement is called total reflectance. Measurements are made, wavelength by wavelength, relative to the reflectance of a calibrated standard of known spectral reflectance. The spectral sensitivity of the detector does not matter so long as it is sufficiently sensitive to allow reliable measurements.

The arrangement of Fig. 10 ensures that all components of diffusely reflected light are equally weighted and that the specular component can be included in or eliminated from the measurement. Often, however, there is no true specular component, but rather a high-intensity lobe with a definite spread. This renders somewhat arbitrary the distinction between the specular and diffuse components. Operationally, the distinction depends on the size of exit port chosen for the specular light trap. Reflectance measurements are usually scaled relative to what a perfectly diffuse, totally reflecting surface would produce if located in the position of the sample. Figure 11 shows the diffuse spectral reflectance curves of a set of enamel paints that are sometimes used as calibration standards.

C. Chromaticity of an Object

The chromaticity of an object depends on the spectral properties of the illuminant as well as those of the object. A quantity $\phi(\lambda)$ is defined as $\rho(\lambda)S(\lambda)$ or $\tau(\lambda)S(\lambda)$, where $\rho(\lambda)$ symbolizes reflectance and $\tau(\lambda)$ symbolizes

TABLE II Color and Color-Rendering Characteristics of Common Light Sources[a]

Test lamp designation	CIE chromaticity coordinates		Correlated color temperature (Kelvins)	CIE general color rendering Index, R_a	CIE special color rendering indices R_i													
	x	y			R_1	R_2	R_3	R_4	R_5	R_6	R_7	R_8	R_9	R_{10}	R_{11}	R_{12}	R_{13}	R_{14}
Fluorescent lamps																		
Warm White	0.436	0.406	3020	52	43	70	90	40	42	55	66	13	−111	31	21	27	48	94
Warm White Deluxe	0.440	0.403	2940	73	72	80	81	71	69	67	83	64	14	49	60	43	73	88
White	0.410	0.398	3450	57	48	72	90	47	49	61	68	20	−104	36	32	38	52	94
Cool White	0.373	0.385	4250	62	52	74	90	54	56	64	74	31	−94	39	42	48	57	93
Cool White Deluxe	0.376	0.368	4050	89	91	91	85	89	90	86	90	88	70	74	88	78	91	90
Daylight	0.316	0.345	6250	74	67	82	92	70	72	78	82	51	−56	59	64	72	71	95
Three-Component A	0.376	0.374	4100	83	98	94	48	89	89	78	88	82	32	46	73	53	95	65
Three-Component B	0.370	0.381	4310	82	84	93	66	65	28	94	83	85	44	69	62	68	90	76
Simulated D_{50}	0.342	0.359	5150	95	93	96	98	95	94	95	98	92	76	91	94	93	94	99
Simulated D_{55}	0.333	0.352	5480	98	99	98	96	99	99	98	98	96	91	95	98	97	98	98
Simulated D_{65}	0.313	0.325	6520	91	93	91	85	91	93	88	90	92	89	76	91	86	92	91
Simulated D_{70}	0.307	0.314	6980	93	97	93	87	92	97	91	91	94	95	82	95	93	94	93
Simulated D_{75}	0.299	0.315	7500	93	93	94	91	93	93	91	94	91	73	83	92	90	93	95
Mercury, clear	0.326	0.390	5710	15	−15	32	59	2	3	7	45	−15	−327	−55	−22	−25	−3	75
Mercury improved color	0.373	0.415	4430	32	10	43	60	20	18	14	60	31	−108	−32	−7	−23	17	77
Metal halide, clear	0.396	0.390	3720	60	52	84	81	54	60	83	59	5	−142	68	55	78	62	88
Xenon, high pressure arc	0.324	0.324	5920	94	94	91	90	96	95	92	95	96	81	81	97	93	92	95
High pressure sodium	0.519	0.418	2100	21	11	65	52	−9	10	55	32	−52	−212	45	−34	32	18	69
Low pressure sodium	0.569	0.421	1740	−44	−68	44	−2	−101	−67	29	−23	−165	−492	20	−128	−21	−39	31
DXW tungsten halogen	0.424	0.399	3190	100	All 100 except for $R_6 = R_{10} = 99$													

[a] Lamps representative of the industry are listed. Variations from manufacturer to manufacturer are likely, especially for the D series of fluorescent lamps and the high-intensity discharge lamps. A high positive value of R_i indicates a small color difference for sample i. A low value of R_i indicates a large color difference.

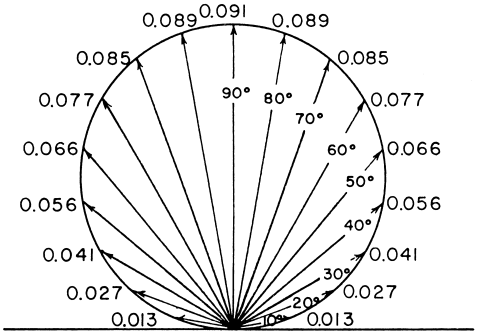


FIGURE 9 Intensity distribution of light reflected from a perfectly diffuse (Lambertian) surface, showing proportion of reflected light within 5° of each indicated direction. [From Boynton, R. M. (1974). *In* "Handbook of Perception" (E. C. Carterette and M. P. Friedman, eds.), Vol. 1. Copyright 1974 Academic Press.]

transmittance. Whereas reflectance ρ is the fraction of incident light returned from a surface, transmittance τ is the fraction of incident light transmitted by an object. Tristimulus values are then calculated as follows:

$$X = k \int \phi_\lambda \bar{x}(\lambda)\, d\lambda,$$

$$Y = k \int \phi_\lambda \bar{y}(\lambda)\, d\lambda,$$

$$Z = k \int \phi_\lambda \bar{z}(\lambda)\, d\lambda.$$

Calculation of chromaticity coordinates then proceeds as described above for sources. If an equal-energy spectrum is assumed, the source term $S(\lambda)$ can be dropped from the definition of $\phi(\lambda)$. When the chromaticity of a surface is specified without specification of the source, an equal-energy spectrum is usually implied.

D. Fluorescence

The practice of colorimetry so far described becomes considerably more complicated if the measured surface exhibits fluorescence. Materials with fluorescing surfaces, when excited by incident light, generally both emit light at a longer wavelength and reflect a portion of the incident light. When making reflectance measurements of nonfluorescent materials, there is no reason to use incident radiation in the ultraviolet, to which the visual mechanism is nearly insensitive. However, for fluorescent materials these incident wavelengths can stimulate substantial radiation at *visible* wavelengths. The full specification of the relative radiance (reflection plus radiation) properties of such surfaces requires the determination, for *each* incident wavelength (including those wavelengths in the ultraviolet known to produce fluorescence), of relative radiance at *all* visible wavelengths, leading to a huge matrix of measurement conditions. As a meaningful and practical shortcut, daylight or a suitable daylight substitute can be used to

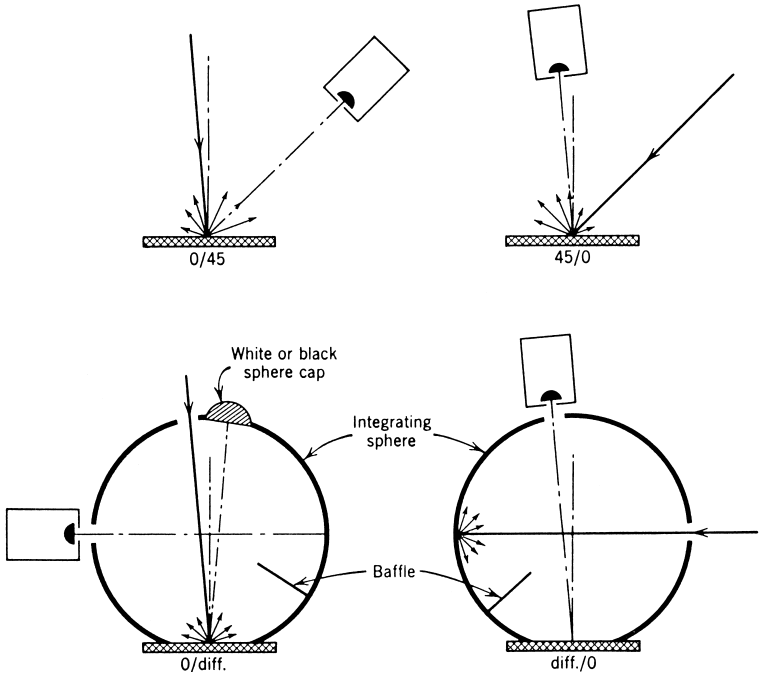


FIGURE 10 Schematic diagram showing the four CIE standard illuminating and viewing geometries for reflectance measurements. [From Wyszecki, G., and Stiles, W. S. (1982). "Color Science: Concepts and Methods. Quantitative Data and Formulae," 2nd ed. Copyright ©1982 John Wiley & Sons, Inc. Reprinted by permission of John Wiley & Sons, Inc.]

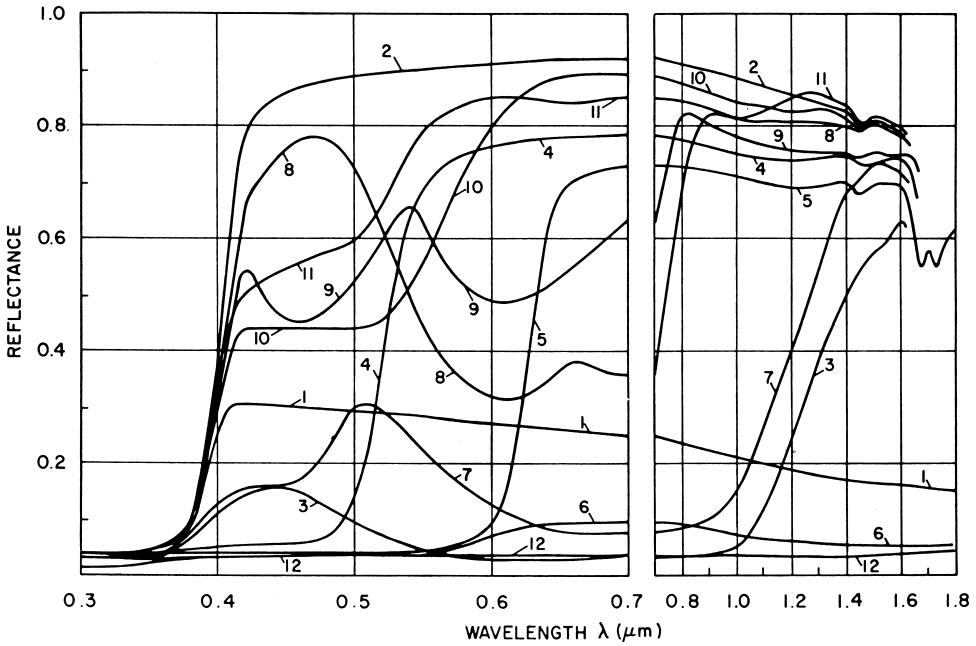


FIGURE 11 Diffuse spectral reflectance curves of a set of enamel paints having the following color appearances: (1) medium gray, (2) white, (3) deep blue, (4) yellow, (5) red, (6) brown, (7) medium green, (8) light blue, (9) light green, (10) peach, (11) ivory, and (12) black. [From Wyszecki, G., and Stiles, W. S. (1982). "Color Science: Concepts and Methods, Quantitative Data and Formulae," 2nd ed. Copyright © 1982 John Wiley & Sons, Inc. Reprinted by permission of John Wiley & Sons, Inc.]

irradiate the sample, and a spectrophotometer can be located at the exit port to register the spectral distribution of the reflected light, which will include the component introduced by fluorescence. Daylight substitutes are so difficult to obtain that the most recent standard illuminant sanctioned by the CIE, called D-65, has been specified only mathematically but has never been perfectly realized. (A suitably filtered, high-pressure xenon arc source comes close.)

E. Optimal Colors

Because of the broadband characteristics of the cone spectral sensitivity functions, most of the spectrum locus in the chromaticity diagram is very well approximated by wave bands as broad as 5 nm. A reflecting surface that completely absorbed all wavelengths of incident broadband (white) light and reflected only a 5-nm wave band would have a chromaticity approximating the midpoint of that wave band along the spectrum locus. Such a surface would also have a very low reflectance because almost all of the incident light would be absorbed. Extending the wave band would increase reflectance, but at the cost of moving the chromaticity inward toward the center of the diagram, with the limit for a nonselective surface being the chromaticity of the illuminant. For any particular reflectance, the domain of possible chromaticities can be calculated; the outer limit of this domain represents the locus of optimal surface colors for that chromaticity.

The all-or-nothing and stepwise reflectance properties required for optimal surface colors do not exist either in nature or in artificially created pigments (see Fig. 11), which tend to exhibit instead gently sloped spectral reflectance functions. For any given reflectance, therefore, the domain of real colors is always much more restricted than the ideal one. Figure 12 shows the CIE chromaticity diagram and the relations between the spectrum locus, the optimal colors of several reflectances, and the real surface colors of the Optical Society of America Uniform Color Scales set.

F. Metamerism Index

As already noted, *metamerism* refers to the phenomenon whereby a color match can occur between stimuli that differ in their spectral distributions. In the domain of reflecting samples the term carries a related, but very different connotation; here the degree of metamerism specifies the tendency of surfaces to change in perceived color, or to resist doing so, as the spectral characteristics of the illuminant are altered. Surfaces greatly exhibiting such changes are said to exhibit a high degree of metamerism, which from a commercial standpoint is undesirable.

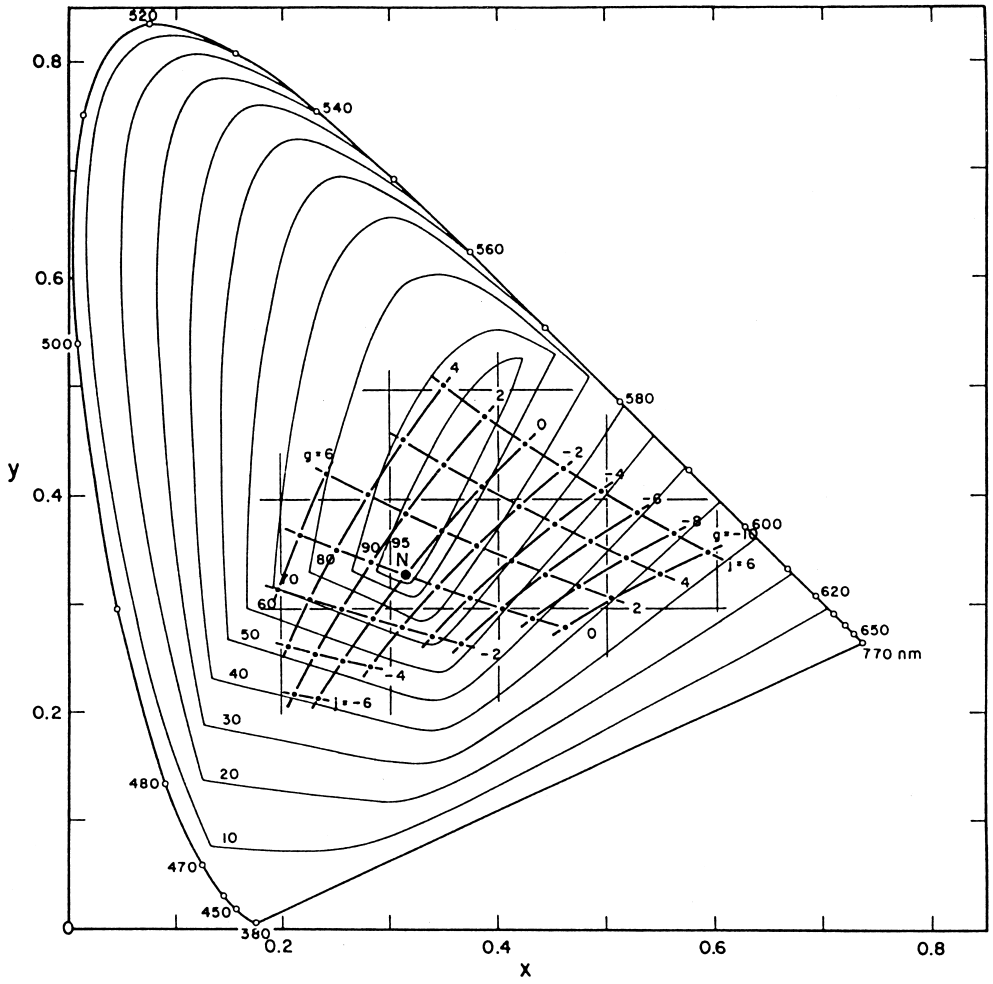


FIGURE 12 Locus of optimal colors of reflectances indicated, also showing the locations of 47 surface colors of ~30% reflectance developed by the Optical Society of America to be equally spaced perceptually. [From Wyszecki, G., and Stiles, W. S. (1982). "Color Science: Concepts and Methods, Quantitative Data and Formulae," 2nd ed. Copyright © 1982 John Wiley & Sons, Inc. Reprinted by permission of John Wiley & Sons, Inc.]

Most indices of metamerism that have been proposed depend either on the assessed change in color of specific surfaces with change in illuminant, calculated by procedures similar to those used for the color-rendering index of illuminants, or on the number of intersections of the spectral reflectance functions of the samples being assessed. For two samples to be metameric, these functions must intersect at least three times; in general, the more intersections, the lower is the degree of metamerism that results, implying more resistance to color change with a change in illuminant. In the limiting case, where the curves are identical, there is no metamerism and the match holds for all illuminants.

Except for monochromatic lights on the curved portion of the spectrum locus in the chromaticity diagram, the number of possible metamers is mathematically infinite. Taking into account the limits of sensitivity of the visual system for the perception of color differences, the number of possible metamers increases as one approaches the white point on the chromaticity diagram, moving inward from the outer limits of realizable reflecting colors.

V. PHYSICAL BASIS OF SURFACE COLOR

The physical basis of the color of a surface is related to processes that alter the spectral distribution of the returned light in the direction of an observer, relative to that of the incident illumination.

A. Color from Organic Molecules

The action of organic molecules, which provide the basis for much of the color seen in nature, has been interpreted

with accelerating precision since about 1950 in the context of molecular orbital theory. The interaction of light with small dye molecules can be completely specified, and although such detailed interpretation remains impractical for large dye molecules, even with currently available supercomputers, the origin of color in organic molecules is considered to be understood in principle at a molecular level.

Most organic dyes contain an extended conjugated chromophore system to which are attached electron donor and electron acceptor groups. Although the wavelength of a "reflected" photon is usually the same as that of the incident one, the reflected photon is not the same particle of light as the incident one. Instead, the surface is more properly regarded as a potential emitter of light, where (in the absence of incandescence, fluorescence, or phosphorescence) incident light is required to trigger the molecular reaction that produces the emitted radiation. Except for fluorescent materials, the number of emitted photons cannot exceed *at any wavelength* the number that are incident, and the frequency of each photon is unchanged, as is its wavelength if the medium is homogeneous. In considering nonfluorescent materials, the subtle exchange of reflected for incident photons is of no practical importance, and the term *reflection* is often used to describe the process as if some percentage of photons were merely bouncing off the surface.

B. Colorants

A colorant is any substance employed to produce reflection that is selective with wavelength. Colorants exist in two broad categories: dyes and pigments. In general, dyes are soluble, whereas pigments require a substrate called a binder. Not all colorants fall into either of these categories. (Exceptions include colorants used in enamels, glasses, and glazes.)

C. Scatter

Because pigments do no exist as dissociated, individual molecules, but instead are bound within particles whose size and distribution may vary, the spectral distribution of the reflected light depends only partly on the reaction of the dye or pigment molecules to absorbed light. In addition to the possibility of being absorbed, reflected, or transmitted, light may also be scattered. Particles that are very small relative to the wavelength of light produce Rayleigh scattering, which varies inversely as the fourth power of wavelength. (Rayleigh scatter causes the sky to appear blue on a clear day; without scatter from atmospheric particles, the sky would be black, as seen from the moon.) As scattering particles become larger, Mie scattering results. Wavelength dependence, which is minimal for large-particle scatter, becomes a factor for particles of intermediate size. The directionality of scattered light is complex and can be compounded by multiple scattering. There are two components of Rayleigh scattering, which are differentially polarized. Mie scattering is even more complicated than the Rayleigh variety, and calculations pertaining to it are possible to carry out only with very large computers.

Scatter also occurs at object surfaces. For example, in "blue-eyed" people and animals, the eye color results mainly from scatter within a lightly pigmented iris. As a powder containing a colorant is ground more finely or is compressed into a solid block, its scattering characteristics change and so does the spectral distribution of the light reflected from it, despite an unchanging molecular configuration of the colorant. Often such scatter is nonselective with wavelength and tends to dilute the selective effects of the colorant. A compressed block of calcium carbonate is interesting in this respect because it comes very close to being a perfectly diffuse, totally reflecting, spectrally nonselective reflector.

D. Other Causes of Spectrally Selective Reflection

The spectral distribution of returned light can also be altered by interference and diffraction. Interference colors are commonly seen in thin films of oil resting on water; digital recording disks now provide a common example of spectral dispersion by diffraction.

Light is often transmitted partially through a material before being scattered or reflected. Various phenomena related to transmitted light per se also give rise to spectrally selective effects. In a transmitting substance, such as glass, light is repeatedly absorbed and reradiated, and in the process its speed is differentially reduced as a function of wavelength. This leads to wavelength-selective refraction and the prismatic dispersion of white light into its spectral components.

The most common colorants in glass are oxides of transition metals. Glass may be regarded as a solid fluid, in the sense of being a disordered, noncrystalline system. The metal oxides enter the molton glass in true solution and maintain that essential character after the glass has cooled and hardened. Whereas much colored glass is used for decorative purposes, color filters for scientific use are deliberately produced with specific densities and spectral distributions caused by selective absorption (which usually also produces some scatter), by reflection from coated surfaces, or by interference.

The visual characteristics of metals result from specular reflection (usually somewhat diffused) which, unlike that from other polished surfaces, is spectrally selective.

If the regular periodicity of their atoms is taken into account, the reflectance characteristics of metals can also be understood in terms of the same molecular orbital theory that applies to organic colorants. In this case it serves as a more fundamental basis for band theory, in terms of which the optical and electrical conductance properties of metals and semiconductors have classically been characterized.

E. Subtractive Color Mixture

The addition of primaries, as described in Section IA, is an example of what is often termed additive color mixture. Four methods of addition were described, all of which have in common the fact that photons of different wavelengths enter the eye from the same, or nearly the same, part of the visual field. There are no significant interactions between photons external to the eye; their integration occurs entirely within the photoreceptors, where photons of different wavelengths are absorbed in separate molecules of photopigments, which for a given photoreceptor are all of the same kind, housed within the cone outer segments.

Subtractive color mixing, on the other hand, is concerned with the modification of spectral light distributions external to the eye by the action of absorptive colorants, which, in the simplest case, can be considered to act in successive layers. Here it is dyes or pigments, not lights, that are mixed. The simplest case, approximated in some color photography processes, consists of layers of nonscattering, selectively absorptive filters. Consider the spectral transmittance functions of the subtractive primaries called cyan and yellow in Fig. 13 and the result of their combination: green. The transmittance function for the resulting green is simply the product, wavelength by wavelength, of the transmittance functions of cyan and yellow. When a third subtractive primary (magenta) is included in the system, blue and red can also be produced by the combinations shown. If all three subtractive primaries are used, very little light can pass through the combination, and the result is black.

If the filters are replaced by dyes in an ideal nonscattering solution, transmittance functions of the cyan, yellow, and magenta primaries can be varied quantitatively, depending on their concentration, with little change of "shape"—that is, each can be multiplied by a constant at each wavelength. By varying the relative concentrations of three dyes, a wide range of colors can be produced, as shown by the line segments on the CIE chromaticity diagram of Fig. 14. Subtractive color mixtures do not fall along straight lines in the chromaticity diagram.

Dichroic filters, such as those used in color television cameras, ideally do not absorb, but instead reflect the component of light not transmitted, so that the two components are complementary in color. By contrast, examination of an ordinary red gelatin filter reveals that the appearance of light reflected from it, as well as that transmitted through it, is red. The explanation for the reflected component is similar to that for colors produced by the application of pigments to a surface.

Consider elemental layers within the filter and a painted surface, each oriented horizontally, with light incident downward. In both cases, the light incident at each elemental layer consists of that not already absorbed or backscattered in the layers above. Within the elemental layer, some fraction of the incident light will be scattered upward, to suffer further absorption and scatter before some of it emerges from the surface at the top. Another fraction will be absorbed within the elemental layer, and the remainder will be transmitted downward, some specularly and some by scatter. In the case of the painted surface, a fraction of the initially incident light will reach the backing. In the case of the red gelatin filter, light emerging at the bottom constitutes the component transmitted through the filter. For the painted surface, the spectral reflectance of the backing will, unless perfectly neutral, alter the spectral distribution of the light reflected upward, with further attenuation and scattering at each elemental layer, until some of the light emerges at the top.

The prediction of color matches involving mixtures of pigments in scattering media is, as the example above suggests, not a simple matter. For this purpose, a theory developed by Kubelka and Munk in 1931, and named after them, is (with variations) most often used. Complex as it is, the theory nevertheless requires so many simplifying assumptions that predictions based on it are only approximate. Sometimes Mie scattering theory is applied to the problem of predicting pigment formulations to match a color specification, but more often empirical methods are used for this purpose.

VI. COLOR DIFFERENCE AND COLOR ORDER

Paradoxically, exact color matches are at the same time very common and very rare. They are common in the sense that any two sections of the same uniformly colored surface or material will usually match physically and therefore also visually. Otherwise, exact color matches are rare. For example, samples of paints of the same name and specification, intended to match, seldom do so exactly if drawn from separate batches. Pigments of identical chemical specification generally do not cause surfaces to match if they are ground to different particle sizes or suspended within different binders. Physical matches of different materials, such as plastics and fabrics, are usually impossible because differing binders or colorants must be used.

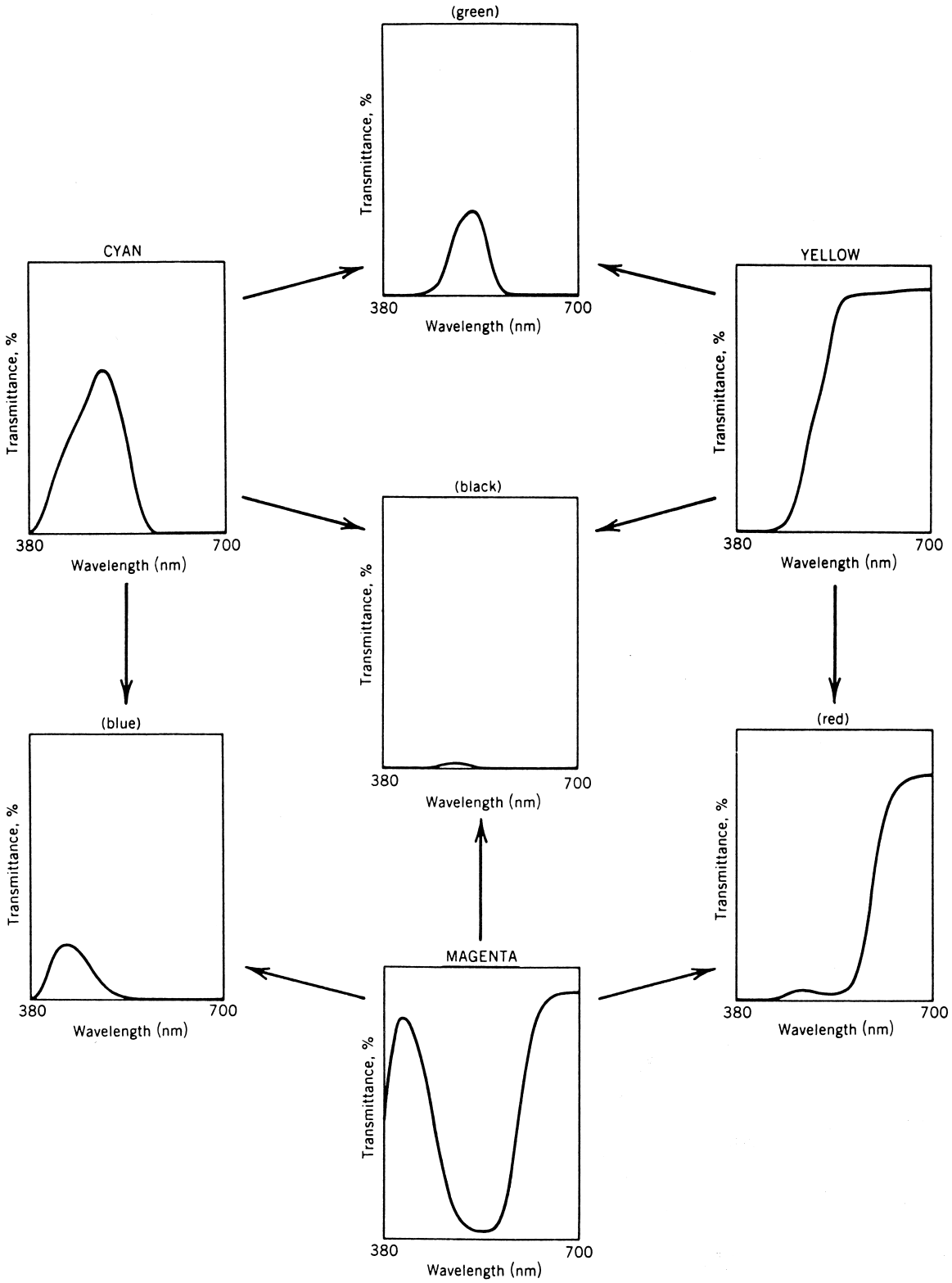


FIGURE 13 Spectrophotometric curves of a set of subtractive primary filters and their mixtures, superimposed in various combinations. [From Billmeyer, F. W., Jr., and Saltzmann, M. (1981). "Principles of Color Technology," 2nd ed. Copyright © 1981 John Wiley & Sons, Inc. Reprinted by permission of John Wiley & Sons, Inc.]

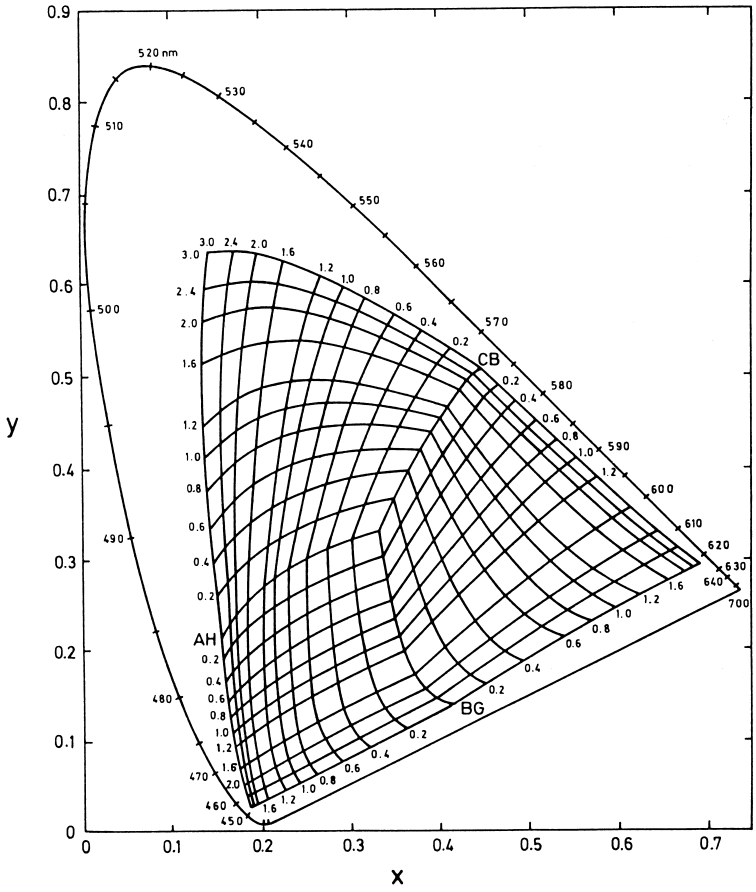


FIGURE 14 Chromaticities in a white light (~6500 K) of indicated combinations of dyes AH, BG, and CB in various concentrations. (Courtesy D. L. MacAdam and Springer-Verlag.)

In such cases—for example, matching a plastic dashboard with the fabric of an automobile seat—metameric matches must suffice; these cannot be perfect for all viewing conditions and observers. Given the difficulty or impossibility of producing perfect matches, it is important to be able to specify tolerances within which imperfect matches will be acceptable.

The issue of color differences on a more global scale will also be considered. Here concern is with the arrangement of colors in a conceptual space that will be helpful for visualizing the relations among colors of all possible kinds—the issue of color order.

A. Color-Difference Data

Figure 15 shows the so-called MacAdam discrimination ellipses plotted in the CIE chromaticity diagram. These were produced more than 40 years ago by an experimental subject who repeatedly attempted to make perfect color matches to samples located at 25 points in chromaticity space. The apparatus provided projected rather than surface colors, but with a specified achromatic surround. For a set of settings at a given reference chromaticity, the apparatus was arranged so that the manipulation of a single control caused chromaticity to change linearly through the physical match point in a specified direction while automatically keeping luminance constant. Many attempted matches were made for each of several directions, as an index of a criterion sensory difference. The standard deviations of which were plotted on both sides of each reference chromaticity. Each of the ellipses of Fig. 15 was fitted to collections of such experimental points. MacAdam developed a system for interpolating between the measured chromaticities, and further research extended the effort to include luminance differences as well, leading to discrimination ellipsoids represented in x–y–Y space. By this criterion of discrimination, there are several million discriminable colors.

Early calculational methods required the use of graphical aids, some of which are still in widespread use for commercial purposes. Very soon it was recognized that, if a formula could be developed for the prediction of

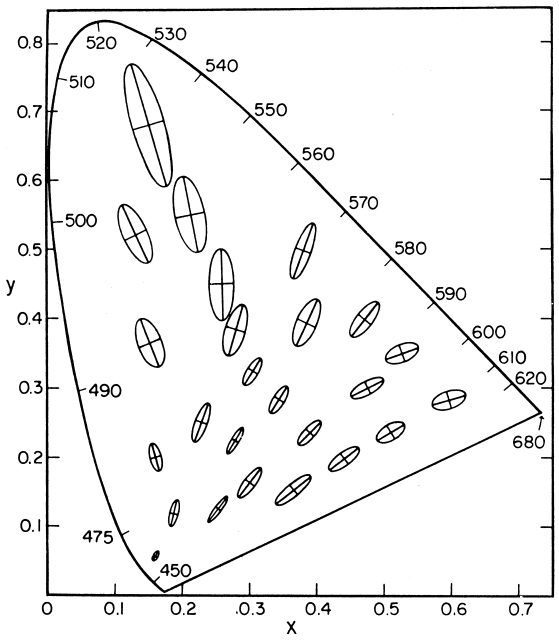


FIGURE 15 MacAdam discrimination ellipses, 10 times actual size. (Courtesy D. L. MacAdam and Springer-Verlag.)

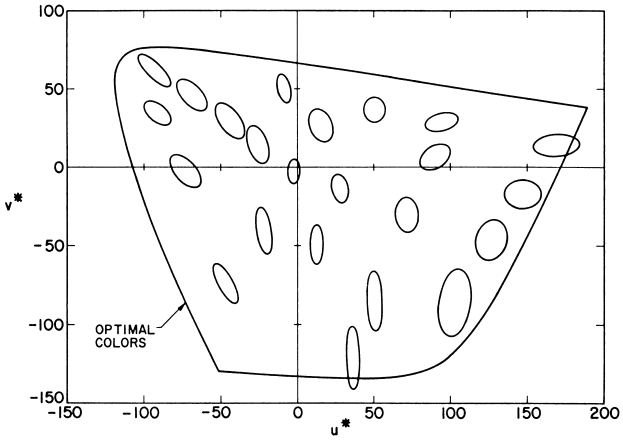


FIGURE 16 MacAdam ellipses ($L^* = 50$) shown in the CIE u^*, v^* diagram. [From Wyszecki, G., and Stiles, W. S. (1982). "Color Science: Concepts and Methods, Quantitative Data and Formulae," 2nd ed. Copyright ©1982 John Wiley & Sons, Inc. Reprinted by permission of John Wiley & Sons, Inc.]

just-discriminable color differences, measurements of color differences could be made with photoelectric colorimeters. Many such formulas have been proposed. To promote uniformity of use, the CIE in 1976 sanctioned two systems called CIELAB and CIELUV, the second of which will be described here.

B. CIELUV Color Difference Formulas

It has long been recognized that the 1931 CIE chromaticity diagram is perceptually nonuniform, as revealed by the different sizes and orientations of the MacAdam ellipses plotted thereon. For the evaluation of chromatic differences, the ideal chromaticity space would be isotropic, and discrimination ellipsoids would everywhere be spheres whose cross sections in a constant-luminance plane would plot as circles of equal size.

Many different projections of the chromaticity diagram are possible; these correspond to changes in the assumed primaries, all of which convey the same basic information about color matches. The projection of Fig. 16 is based on the following equations:

$$u' = 4X/(X + 15Y + 3Z);$$

$$v' = 9Y/(X + 15Y + 3Z).$$

The CIELUV formula is based on the chromatic scaling defined by this transformation combined with a scale of lightness. The system is related to a white reference object having tristimulus values X_n, Y_n, Z_n, with Y_n, usually taken as 100; these are taken to be the tristimulus values of a perfectly reflecting diffuser under a specified white illuminant.

There quantities are then defined as

$$L^* = 116Y/Y_n - 16,$$

$$u^* = 13L^*(u' - u'_n),$$

$$v^* = 13L^*(v' - v'_n).$$

These attempt to define an isotropic three-dimensional space having axes L^*, u^*, and v^*, such that a color difference ΔE^*_{uv} is defined as

$$\Delta E^*_{uv} = \Delta L^* + \Delta u^* + \Delta v^*.$$

The MacAdam ellipses, plotted on the u^*, v^* diagram, are more uniform in size and orientation than in the CIE diagram. Without recourse to nonlinear transformations, this is about the greatest degree of uniformity possible. These data and the CIE's color difference equations are recommended for use under conditions in which the observer is assumed to be adapted to average daylight; they are not recommended by the CIE for other conditions of adaptation.

It is not difficult to write a computer program that will calculate the ΔE^*_{uv} values appropriate to each member of a pair of physical samples. Starting with knowledge of the spectral reflectance distributions of the samples and the spectral irradiance distribution of the illuminant, one calculates the tristimulus values X, Y, and Z. From these, the L^*, u^*, and v^* values for each sample are calculated and inserted into the final formula. Given that voltages proportional to tristimulus values can be approximated using suitably filtered photocells and electronics, it is a

short step to the development of fully automated devices that, when aimed in turn at each of two surfaces, will register a color difference value.

The CIELUV formula is only one of more than a dozen schemes that have been suggested for calculating color differences, some simpler but most more elaborate. None of these performs as well as would be desired. Correlations of direct visual tests with predictions made by the best of these systems, including CIELUV, account for only about half the experimental variance. Different formulas make predictions that correlate no better than this with one another. In using CIELUV to predict color differences in self-luminous displays, agreement is lacking concerning the appropriate choice of reference white. For industrial applications in which reflecting materials are being evaluated, differential weighting of the three components entering into the CIELUV equation may be helpful, and to meet the demands of specific situations, doing so can significantly improve the predictive power of the system. For example, when samples of fabrics are being compared, tolerance for luminance mismatches tends to be greater than for mismatches along the chromatic dimensions.

Despite these problems and limitations, calculations of chromatic differences by formula has proved useful, automated colorimeters for doing so exist, and the practice may be regarded as established, perhaps more firmly than it should be. Room for improvement at a practical level and for a better theoretical understanding of the problem certainly exists.

C. Arrangement of Colors

In the years before the development of the CIE system of color specification, which is based on radiometric measurement, colors could be described only by appeal to labeled physical samples. Any two people possessing a common collection of samples could then specify a color by reference to its label. Whereas in principle such sets of colors could be randomly arranged, an orderly arrangement is clearly preferable in which adjacent colors differ by only a small amount.

For more than 100 years, it has been thought that colors can be continuously arranged in a domain composed of two cones sharing a common base, as shown in Fig. 17. A line connecting the cone apices defines the axis of achromatic colors, ranging from white at the top to black at the bottom. A horizontal plane intersecting one of the cones, or their common base, defines a set of colors of equal lightness. Within such a plane colors can be represented in an orderly way, with gray at the center and colors of maximum saturation on the circumference. The hues on the circumference are arranged as they are in the spectrum, in the order red, orange, yellow, green, blue, and violet, with

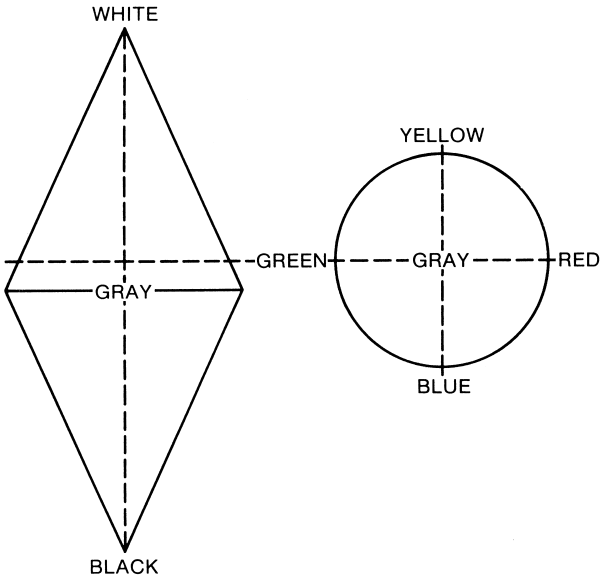


FIGURE 17 Representation of three-dimensional domain of surface colors.

the addition of a range of purples shading back to red. In moving from gray at the center toward a saturated hue, the hue remains constant while the white content of the color gradually diminishes as its chromatic content increases, all at constant lightness. As the intersecting horizontal plane is moved upward, the represented colors are more highly reflecting and correspondingly lighter in appearance, but their gamut is more restricted as must be so if only white is seen at the top. As the intersecting horizontal plane is moved down ward, colors become less reflecting and darker in appearance, with progressively more restricted gamuts, until at the bottom only a pure black is seen.

Considering a typical cross section, a chromatic gray is at its center. Significant features, first observed by Newton, pertain to the circumferential colors. First, they represent the most highly saturated colors conceivable at that level of lightness. Second, adjacent hues, if additively mixed, form legitimate blends; for example, there is a continuous range of blue-greens between blue and green. Third, opposite hues cannot blend. For example, yellow and blue, when additively mixed, produce a white that contains no trace of either component, and the sensations of yellow and blue are never simultaneously experienced in the same spatial location. Two colors that when additively mixed yield a white are called complementary colors, and in this kind of color-order system they plot on opposite sides of the hue circle.

There are several such systems in common use, of which only one, the Munsell system, will be described here. In this system, the vertical lightness axis is said to vary in *value* (equivalent to lightness) from 0 (black) to 10 (white). At any given value level, colors are arranged as described

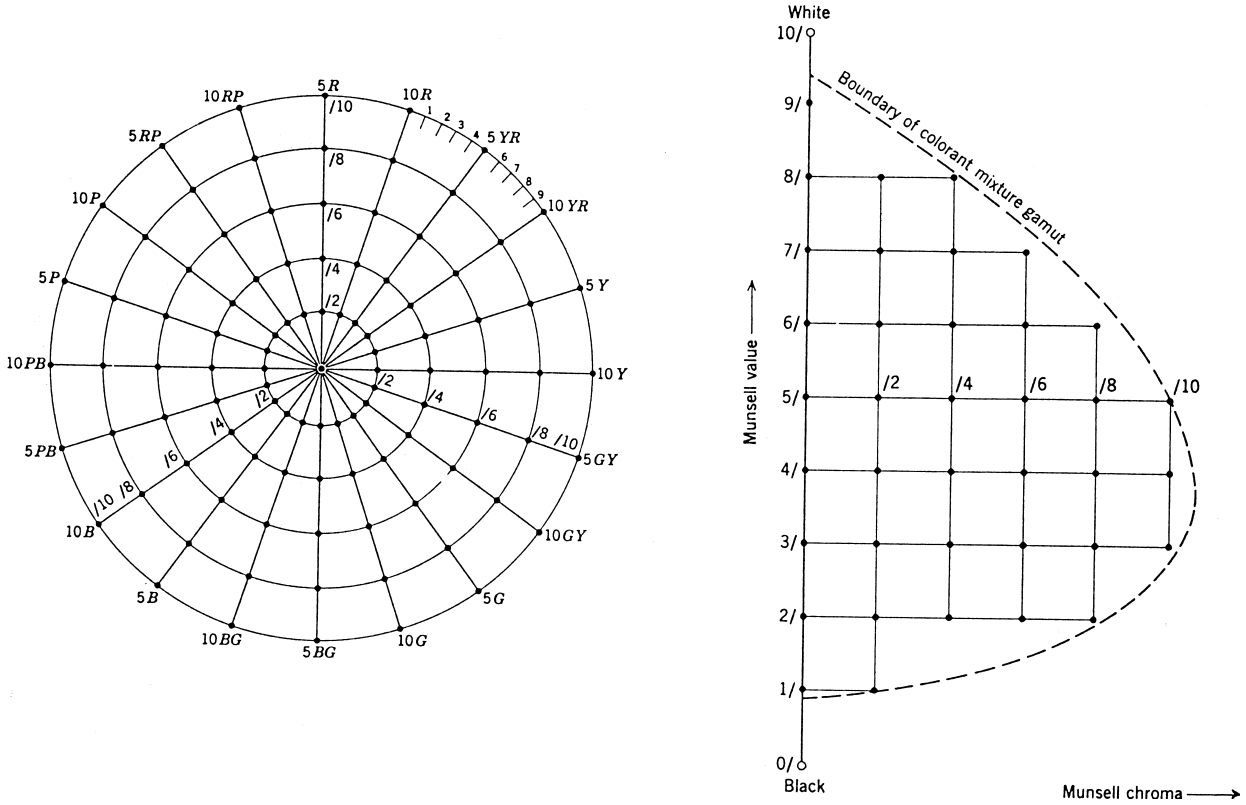


FIGURE 18 Organization of colors in the Munsell system. [From Wyszecki, G., and Stiles, W. S. (1982). "Color Science: Concepts and Methods, Quantitative Data and Formulae," 2nd ed. Copyright ©1982 John Wiley & Sons, Inc. Reprinted by permission of John Wiley & Sons, Inc.]

above but labeled circumferentially according to the hues blue, green, yellow, red, purple (and adjacent blends), and radially according to their saturation (called chroma). Figure 18 illustrates the system.

CIE chromaticity coordinates and Y values have been determined for each of the Munsell samples—the so-called Munsell renotations. A rubber-sheet type of transformation exists between the locations of Munsell colors at a given lightness level and their locations the CIE diagram, as shown in Fig. 19, for Munsell value 5. This figure illustrates that color order can also be visualized on the CIE diagram. A limitation of the CIE diagram for this purpose, in addition to its perceptually nonuniform property, is that it refers to no particular lightness level. The dark surface colors, including brown and black, do not exist in isolated patches of light. These colors are seen only in relation to a lighter surround; in general, surface colors are seen in a complex context of surrounding colors and are sometimes called related colors for this reason.

Although surrounding colors can profoundly influence the appearance of a test color in the laboratory situation, these effects are seldom obvious in natural environments and probably represent an influence of the same processes responsible for color constancy. This concept refers to the fact that colors appear to change remarkably little despite changes in the illuminant that materially alter the spectral distribution of the light reaching the retina. In other words, the perceived color of an object tends, very adaptively, to be correlated with its relative spectral reflectance, so that within limits color seems to be an unchanging characteristic of the object rather than triply dependent, as it actually is, on the characteristics of the illuminant, object, and observer.

VII. PHYSIOLOGICAL BASIS OF COLOR VISION

Progress in modern neuroscience, including a progressively better understanding of sensory systems in physical and chemical terms, has been especially rapid since about 1950 and continues to accelerate. The following is a very brief summary of some highlights and areas of ignorance related to color vision.

The optical system of the eye receives light reflected from external objects and images it on the retina, with a

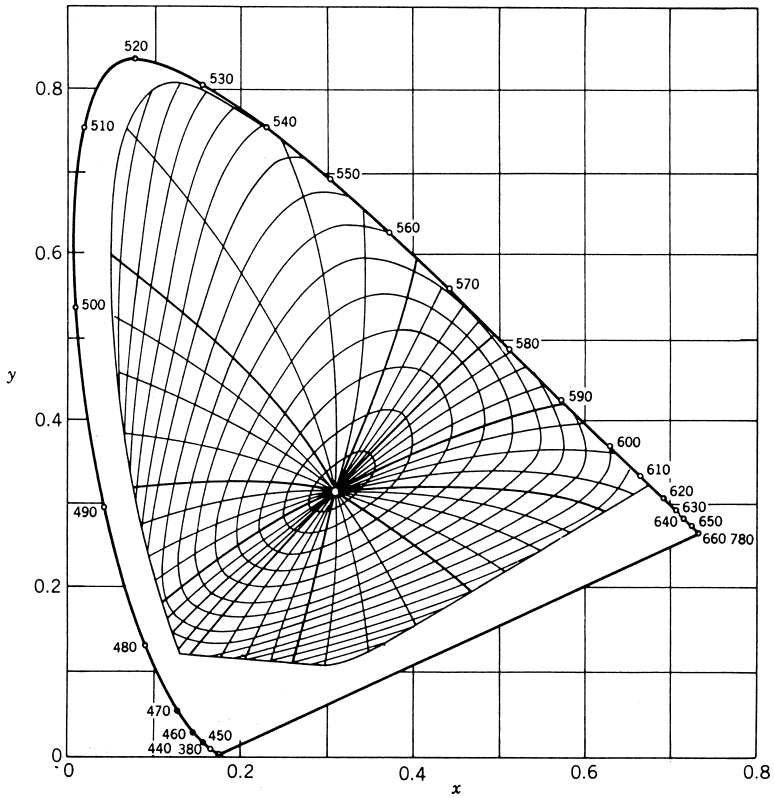


FIGURE 19 Location of circles of constant chroma and lines of constant hue from the Munsell system plotted on the CIE chromaticity diagram. [From Billmeyer, F. W., Jr., and Saltzmann, M. (1981). "Principles of Color Technology," 2nd ed. Copyright © 1981 John Wiley & Sons, Inc. Reprinted by permission of John Wiley & Sons, Inc.]

spectral distribution that is altered by absorption in the eye media. By movements of the eyes that are ordinarily unconsciously programmed, the images of objects of interest are brought to the very center of a specialized region of the retina, known as the fovea centralis, where we enjoy our most detailed spatial vision. The color of objects in the peripheral visual field plays an important role in this process.

The spectral sensitivities of the three classes of cone photoreceptors shown in Fig. 2 depend on the action spectra of three classes of photopigments, each uniquely housed in one type of cone. The cones are very nonuniformly distributed in the retina, being present in highest density in the fovea and falling off rapidly to lower density levels across the remainder of the retina.

The colors of small stimuli seen in the periphery are not registered so clearly as in foveal vision, but if fields are enlarged sufficiently, the periphery is capable of conveying a great deal of information about color. In the fovea there are few if any short-wavelength-sensitive (S) cones, and the long-wavelength-sensitive (L) and middle-wavelength-sensitive (M) cones are present in roughly equal numbers and very high density.

The L and M cones subserve spatial vision and also provide chromatic information concerned with the balance between red and green. Outside the fovea, the proportion of S cones increases but is always very small. The coarseness of the S-cone mosaic makes it impossible for them to contribute very much to detailed spatial vision; instead, they provide information concerned almost exclusively with the second dimension of chromatic vision.

As noted earlier, color is coded initially in terms of the ratios of excitation of the three kinds of cones. A single cone class in isolation is color-blind, because any two spectral distributions can excite such cones equally if appropriate relative intensities are used. The same is true of the much more numerous rod photoreceptors, which can lead to total color-blindness in night vision, where the amount of light available is often insufficient to be effective for cones. At intermediate radiance levels, rods influence both color appearance and, to a degree, color matches. Interestingly, vision at these levels remains trichromatic in the sense that color matches can be made using three primaries and three controls. This suggests that rods feed their signals into pathways shared by cones, a fact documented by direct electrophysiological experiment.

Light absorption in the cones generates an electrical signal in each receptor and modulates the rate of release of a

neurotransmitter at the cone pedicles, where they synapse with horizontal and bipolar cells in the retina. The latter deliver their signals to the ganglion cells, whose long, slender axons leave the eye at the optic disk as a sheathed bundle, the optic nerve. Within this nerve are the patterns of impulses, distributed in about a million fibers from each eye, by means of which the brain is exclusively informed about the interaction of light with objects in the external world, on the basis of which form and color are perceived.

Lateral interactions are especially important for color vision. The color of a particular area of the visual scene depends not only on the spectral distribution of the light coming from that area, but also on the spectral distribution (and quantity) of light coming from other regions of the visual field. In both retina and brain, elaborate laternal interconnections are sufficient to provide the basis for these interactions.

Figure 20 summarizes, in simplified fashion, a current model of retinal function. The initial trichromacy represented by L, M, and S cones is transformed within the retina to a different trichromatic code. The outputs of the L and M cones are summed to provide a luminance signal, which is equivalent to the quantity of light as defined by flicker photometry. The L and M cone outputs are also differenced to form a red—green signal, which carries information about the relative excitations of the L and M cones. An external object that reflects long-wavelength light selectively excites the L cones more than the M, causing the red–green difference signal to swing in the red direction. A yellow–blue signal is derived from the difference between the luminance signal and that from the S cones.

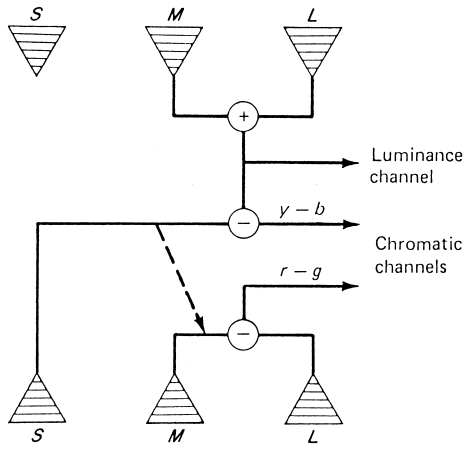


FIGURE 20 Opponent-color model of human color vision at the retinal level.

The color appearance of isolated fields of color, whether monochromatic or spectrally mixed, can be reasonably well understood in terms of the relative strengths of the red–green, yellow–blue, and luminance signals as these are affected in turn by the strength of the initial signals generated in the three types of cones. In addition, increasing S-cone excitation appears to influence the red–green opponent color signal in the red direction.

There has been a great deal of investigation of the anatomy and physiology of the brain as it relates to color perception, based on an array of techniques that continues to expand. The primary visual input arrives in the striate cortex at the back of the head; in addition, there are several other brain centers that receive visual input, some of which seem specifically concerned with chromatic vision. The meaning of this activity for visual perception is not yet clear. In particular, it is not yet known exactly which kinds or patterns of activity immediately underly the sensation of color or exactly where they are located in the brain.

SEE ALSO THE FOLLOWING ARTICLES

BONDING AND STRUCTURE IN SOLIDS • COATINGS, COLORANTS, AND PAINTS • GLASS • HOLOGRAPHY • LIGHT SOURCES • OPTICAL DIFFRACTION • RADIOMETRY AND PHOTOMETRY • SCATTERING AND RECOILING SPECTROSCOPY

BIBLIOGRAPHY

Billmeyer, F. W., Jr., and Saltzman, M. (1981). "Principles of Color Technology," 2nd ed., Wiley, New York.
Boynton, R. M. (1979). "Human Color Vision," Holt, New York.
Grum, F., and Bartleson, C. J. (1980). "Optical Radiation Measurements," Vol. 2, Academic Press, New York.
Kaufman, J. E., ed. (1981). "IES Lighting Handbook: Reference Volume," Illuminating Engineering Society of North America, New York.
MacAdam, D. L. (1981). "Color Measurement: Theme and Variations," Springer-Verlag, Berlin and New York.
Marmion, D. M. (1991). "Handbook of U.S. Colorants," 3rd ed., Wiley, New York.
Mollon, J. D., and Sharpe, L. T., eds. (1983). "Colour Vision: Physiology and Psychophysics," Academic Press, New York.
Nassau, K. (1983). "The Physics and Chemistry of Color," Wiley, New York.
Wyszecki, G., and Stiles, W. S. (1982). "Color Science: Concepts and Methods, Quantitative Data and Formulae," 2nd ed, Wiley, New York.
Zrenner, E. (1983). "Neurophysiological Aspects of Color Vision in Primates," Springer-Verlag, Berlin and New York.

Critical Data in Physics and Chemistry

David R. Lide, Jr.
National Institute of Standards and Technology (Retired)

Bettijoyce B. Lide
National Institute of Standards and Technology

GLOSSARY

Data Factual information, usually expressed in numerical form, that is derived from an experiment, observation, or calculation.

Database An organized collection of data; the term generally implies that the data are expressed in a computer-readable form.

Evaluated data Data whose accuracy has been assessed through an independent review.

THE TERM CRITICAL DATA refers to measured properties of well-defined substances or materials that have been carefully evaluated and organized for convenient use by scientists and engineers. Such collections of data have traditionally been published as handbooks or tables, which have served as basic reference sources for the technical community. Modern computer technology makes it possible to express these collections as databases, which can be stored, retrieved, and accessed in a variety of ways.

I. HISTORY OF CRITICAL DATA PROGRAMS

As physics and chemistry developed into active scientific disciplines in the eighteenth and nineteenth centuries, it was recognized that the numerical results of experiments and observations were valuable to other researchers, often many years after the data were initially obtained. The archival research literature began to serve the function of a

storage medium for these data. By the end of the nineteenth century, this literature had grown to the point that locating previously published data was time consuming and difficult. This led to the practice of compiling data from the primary literature and publishing this information in handbook format. An early example was the Landolt-Börnstein tables, *Numerical Data and Functional Relationships in Science and Technology*, which first appeared in 1883. Scientists came to depend on such tables and handbooks for quick access to data on physical and chemical properties.

The process of compiling data from the literature often revealed inconsistencies and discrepancies, which indicated errors in the original research. Thus, it became evident that some form of critical selection or evaluation of the published data was highly desirable. The first broad-coverage handbook to attempt this was the *International Critical Tables*, a seven-volume set of data books published in the 1920s. Experts from many nations evaluated the available data in their specialty areas and selected recommended values for the final publication. Further efforts of this nature were started in the 1930s and 1940s in such important areas of physical science as thermodynamics and atomic physics. In the 1950s, programs for the collection and evaluation of data in nuclear physics were established at Brookhaven and Oak Ridge National Laboratories. As scientific research has expanded and the technological applications of research findings have increased, it has become more and more evident that a critically evaluated base of physical and chemical data is essential for the orderly progress of science and technology.

II. ROLE OF THE NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY

Scientists from the U.S. National Bureau of Standards (NBS), whose name was changed to the National Institute of Standards and Technology (NIST) in 1988, played a prominent part in these early critical data projects. In the mid-1960s, NBS was designated as the national focal point for such activities in the United States. It undertook the coordination of a set of activities, known as the National Standard Reference Data System (NSRDS), conducted at universities, industrial laboratories, and NIST itself. Some of these projects were long-term efforts, referred to as data centers, in which relevant data were systematically compiled from the scientific literature, evaluated, and organized into databases. Examples of such data centers include the Atomic Energy Levels Data Center and the Crystal Data Center at NIST and the Radiation Chemistry Data Center at Notre Dame University.

Other organizations have also been active in data compilation and evaluation. Such federal agencies as the Department of Energy, Department of Defense, and National Aeronautics and Space Agency have supported selected data projects relevant to their missions. Certain industrial trade associations (e.g., Gas Producers Association and International Copper Research Association) have sponsored data compilation projects of interest to the industry in question. Many professional societies take an active role in sponsoring or coordinating data projects. Examples include the American Institute of Chemical Engineers (Design Institute for Physical Property Data), ASM International (Alloy Phase Diagram Program), and American Society of Mechanical Engineers (steam properties and other data). The National Academy of Sciences–National Research Council has helped to assess needs for data and has examined several national issues associated with access by scientists to data needed in their research.

III. INTERNATIONAL ACTIVITIES

Like many other aspects of science, data compilation efforts can be carried out more efficiently if there is cooperation at an international level. This is particularly important when physical and chemical data affect technological issues, such as performance specifications for articles in international trade or rules for custody transfer of commodities. An early example of the need for international agreement on physical data is provided by the International Association for the Properties of Steam (IAPS). This group was established more than 60 years ago with the aim of reaching international agreement on the thermophysical properties of water and steam, which are crucial in specifying the performance of turbines, boilers, and pumps. Its international steam tables have been adopted as standards for trade and commerce, as well as for scientific applications.

Several international scientific unions have played a role in data compilation. In particular, the International Union of Pure and Applied Chemistry sponsors projects that deal with various types of chemical data. Unions in the geosciences are concerned with data such as terrestrial magnetism, where correlations with geographic location are important. There are also intergovernmental organizations such as the International Atomic Energy Agency (IAEA), which has evaluated data from nuclear and atomic physics that are important in energy research and development.

In 1966, the International Council of Scientific Unions established a standing Committee on Data for Science and Technology (CODATA), with a mandate to improve the quality, reliability, processing, management, and accessibility of data of importance to science and technology. CODATA has representation from the major countries and scientific unions and approaches data issues on both an

TABLE I CODATA 1998 Recommended Values of the Fundamental Physical Constants

Quantity	Symbol	Value	Unit	Relative std. uncert. u_r
Speed of light in vacuum	c, c_0	299 792 458	$m\,s^{-1}$	(exact)
Magnetic constant	μ_0	$4\pi \times 10^{-7}$ $= 12.566\,370\,614\ldots \times 10^{-7}$	$N\,A^{-2}$ $N\,A^{-2}$	(exact)
Electric constant $1/\mu_0 c^2$	ε_0	$8.854\,187\,817\ldots \times 10^{-12}$	$F\,m^{-1}$	(exact)
Newtonian constant of gravitation	G	$6.673(10) \times 10^{-11}$	$m^3\,kg^{-1}\,s^{-2}$	1.5×10^{-3}
Planck constant $h/2\pi$	h	$6.626\,068\,76(52) \times 10^{-34}$	$J\,s$	7.8×10^{-8}
	$\hbar$	$1.054\,571\,596(82) \times 10^{-34}$	$J\,s$	7.8×10^{-8}
Elementary charge	e	$1.602\,176\,462(63) \times 10^{-19}$	C	3.9×10^{-8}
Magnetic flux quantum $h/2e$	Φ_0	$2.067\,833\,636(81) \times 10^{-15}$	Wb	3.9×10^{-8}
Conductance quantum $2e^2/h$	G_0	$7.748\,091\,696(28) \times 10^{-5}$	S	3.7×10^{-9}
Electron mass	m_e	$9.109\,381\,88(72) \times 10^{-31}$	kg	7.9×10^{-8}
Proton mass	m_p	$1.672\,621\,58(13) \times 10^{-27}$	kg	7.9×10^{-8}
Proton-electron mass ratio	m_p/m_e	1 836.152 6675(39)		2.1×10^{-9}
Fine-structure constant $e^2/4\pi\epsilon_0\hbar c$	α	$7.297\,352\,533(27) \times 10^{-3}$		3.7×10^{-9}
Inverse fine-structure constant	α^{-1}	137.035 999 76(50)		3.7×10^{-9}
Rydberg constant $\alpha^2 m_e c/2h$	R_∞	10 973 731.568 548(83)	m^{-1}	7.6×10^{-12}
Avogadro constant	N_A, L	$6.022\,141\,99(47) \times 10^{23}$	mol^{-1}	7.9×10^{-8}
Faraday constant $N_A e$	F	96 485.3415(39)	$C\,mol^{-1}$	4.0×10^{-8}
Molar gas constant	R	8.314 472(15)	$J\,mol^{-1}\,K^{-1}$	1.7×10^{-6}
Boltzmann constant R/N_A	k	$1.380\,650\,3(24) \times 10^{-23}$	$J\,K^{-1}$	1.7×10^{-6}
Stefan-Boltzmann constant $(\pi^2/60)k^4/\hbar^3 c^2$	σ	$5.670\,400(40) \times 10^{-8}$	$W\,m^{-2}\,K^{-4}$	7.0×10^{-6}
Non-SI units accepted for use with the SI				
Electron volt: (e/C) J	eV	$1.602\,176\,462(63) \times 10^{-19}$	J	3.9×10^{-8}
(Unified) atomic mass unit $1u = m_u = \frac{1}{12}m(^{12}C)$ $= 10^{-3}\,kg\,mol^{-1}/N_A$	u	$1.660\,538\,73(13) \times 10^{-27}$	kg	7.9×10^{-8}

Source: Mohr, P. J., and Taylor, B. N., *J. Phys. Chem. Ref. Data*, in press.

international and an interdisciplinary basis. It has provided recommended sets of certain key values, such as fundamental physical constants and important thermodynamic properties, which have been generally accepted for international use. CODATA also serves as a forum for reaching consensus on standards and formats for presenting data, and it carries out several educational activities such as conferences, training courses, and preparation of tutorial publications.

The best current values of some frequently used physical and chemical data published by various data evaluation groups are presented in Tables I to VIII.

IV. METHODS OF EVALUATING DATA

The question of how to evaluate published data is not easy to answer in a general way. Specific methodologies have been developed in some fields, and these have certain elements in common. However, a technique effective for one physical property may be entirely unsuitable for another.

A common feature of most evaluation efforts is the reduction of all published data to the same basis. Corrections for changes in temperature scale, atomic weights, fundamental constants, conversion relations, and other factors must be made before a true evaluation can be started. This often requires considerable effort to deduce the subsidiary data used by the original authors.

Critical evaluation implies a process of independent assessment of the reliability of data appearing in the literature. This process should be conducted by scientists who are familiar with the type of data in question and who have had experience in the measurement techniques that produced the data. There are usually some subjective elements of the evaluation process. For example, the evaluator will generally have a feeling for the accuracy of each measurement technique and for the pitfalls that can lead to unsuspected errors. The reputation of the researcher or

TABLE II IUPAC Atomic Weights (1995)

Atomic number	Name	Symbol	Atomic weight	Atomic number	Name	Symbol	Atomic weight
1	Hydrogen	H	1.00794(7)	56	Barium	Ba	137.327(7)
2	Helium	He	4.002602(2)	57	Lanthanum	La	138.9055(2)
3	Lithium	Li	6.941(2)	58	Cerium	Ce	140.116(1)
4	Beryllium	Be	9.012182(3)	59	Praseodymium	Pr	140.90765(2)
5	Boron	B	10.811(7)	60	Neodymium	Nd	144.24(3)
6	Carbon	C	12.0107(8)	61	Promethium	Pm	[145]
7	Nitrogen	N	14.00674(7)	62	Samarium	Sm	150.36(3)
8	Oxygen	O	15.9994(3)	63	Europium	Eu	151.964(1)
9	Fluorine	F	18.9984032(5)	64	Gadolinium	Gd	157.25(3)
10	Neon	Ne	20.1797(6)	65	Terbium	Tb	158.92534(2)
11	Sodium	Na	22.989770(2)	66	Dysprosium	Dy	162.50(3)
12	Magnesium	Mg	24.3050(6)	67	Holmium	Ho	164.93032(2)
13	Aluminum	Al	26.981538(2)	68	Erbium	Er	167.26(3)
14	Silicon	Si	28.0855(3)	69	Thulium	Tm	168.93421(2)
15	Phosphorus	P	30.973761(2)	70	Ytterbium	Yb	173.04(3)
16	Sulfur	S	32.066(6)	71	Lutetium	Lu	174.967(1)
17	Chlorine	Cl	35.4527(9)	72	Hafnium	Hf	178.49(2)
18	Argon	Ar	39.948(1)	73	Tantalum	Ta	180.9479(1)
19	Potassium	K	39.0983(1)	74	Tungsten	W	183.84(1)
20	Calcium	Ca	40.078(4)	75	Rhenium	Re	186.207(1)
21	Scandium	Sc	44.955910(8)	76	Osmium	Os	190.23(3)
22	Titanium	Ti	47.867(1)	77	Iridium	Ir	192.217(3)
23	Vanadium	V	50.9415(1)	78	Platinum	Pt	195.078(2)
24	Chromium	Cr	51.9961(6)	79	Gold	Au	196.96655(2)
25	Manganese	Mn	54.938049(9)	80	Mercury	Hg	200.59(2)
26	Iron	Fe	55.845(2)	81	Thallium	Tl	204.3833(2)
27	Cobalt	Co	58.933200(9)	82	Lead	Pb	207.2(1)
28	Nickel	Ni	58.6934(2)	83	Bismuth	Bi	208.98038(2)
29	Copper	Cu	63.546(3)	84	Polonium	Po	[209]
30	Zinc	Zn	65.39(2)	85	Astatine	At	[210]
31	Gallium	Ga	69.723(1)	86	Radon	Rn	[222]
32	Germanium	Ge	72.61(2)	87	Francium	Fr	[223]
33	Arsenic	As	74.92160(2)	88	Radium	Ra	[226]
34	Selenium	Se	78.96(3)	89	Actinium	Ac	[227]
35	Bromine	Br	79.904(1)	90	Thorium	Th	232.0381(1)
36	Krypton	Kr	83.80(1)	91	Protactinium	Pa	231.03588(2)
37	Rubidium	Rb	85.4678(3)	92	Uranium	U	238.0289(1)
38	Strontium	Sr	87.62(1)	93	Neptunium	Np	[237]
39	Yttrium	Y	88.90585(2)	94	Plutonium	Pu	[244]
40	Zirconium	Zr	91.224(2)	95	Americium	Am	[243]
41	Niobium	Nb	92.90638(2)	96	Curium	Cm	[247]
42	Molybdenum	Mo	95.94(1)	97	Berkelium	Bk	[247]
43	Technetium	Tc	[98]	98	Californium	Cf	[251]
44	Ruthenium	Ru	101.07(2)	99	Einsteinium	Es	[252]
45	Rhodium	Rh	102.90550(2)	100	Fermium	Fm	[257]
46	Palladium	Pd	106.42(1)	101	Mendelevium	Md	[258]
47	Silver	Ag	107.8682(2)	102	Nobelium	No	[259]
48	Cadmium	Cd	112.411(8)	103	Lawrencium	Lr	[262]
49	Indium	In	114.818(3)	104	Rutherfordium	Rf	[261]
50	Tin	Sn	118.710(7)	105	Dubnium	Db	[262]
51	Antimony	Sb	121.760(1)	106	Seaborgium	Sg	[266]
52	Tellurium	Te	127.60(3)	107	Bohrium	Bh	[264]
53	Iodine	I	126.90447(3)	108	Hassium	Hs	[269]
54	Xenon	Xe	131.29(2)	109	Meitnerium	Mt	[268]
55	Cesium	Cs	132.90545(2)				

Note: Numbers in parentheses represent the uncertainty in the last digit. Values in brackets are the mass numbers of the longest-lived isotope of elements for which a standard atomic weight cannot be defined.

Source: Pure Appl. Chem. **68**, 2339 (1996).

TABLE III Ground Levels and Ionization Energies for the Neutral Atoms

Z	Element		Ground-state configuration	Ground level	Ionization energy (eV)
1	H	Hydrogen	$1s$	$^2S_{1/2}$	13.5984
2	He	Helium	$1s^2$	$^1S_0$	24.5874
3	Li	Lithium	$1s^2\ 2s$	$^2S_{1/2}$	5.3917
4	Be	Beryllium	$1s^2\ 2s^2$	$^1S_0$	9.3227
5	B	Boron	$1s^2\ 2s^2\ 2p$	$^2P^o_{1/2}$	8.2980
6	C	Carbon	$1s^2\ 2s^2\ 2p^2$	$^3P_0$	11.2603
7	N	Nitrogen	$1s^2\ 2s^2\ 2p^3$	$^4S^o_{3/2}$	14.5341
8	O	Oxygen	$1s^2\ 2s^2\ 2p^4$	$^3P_2$	13.6181
9	F	Fluorine	$1s^2\ 2s^2\ 2p^5$	$^2P^o_{3/2}$	17.4228
10	Ne	Neon	$1s^2\ 2s^2\ 2p^6$	$^1S_0$	21.5646
11	Na	Sodium	[Ne] $3s$	$^2S_{1/2}$	5.1391
12	Mg	Magnesium	[Ne] $3s^2$	$^1S_0$	7.6462
13	Al	Aluminum	[Ne] $3s^2\ 3p$	$^2P^o_{1/2}$	5.9858
14	Si	Silicon	[Ne] $3s^2\ 3p^2$	$^3P_0$	8.1517
15	P	Phosphorus	[Ne] $3s^2\ 3p^3$	$^4S^o_{3/2}$	10.4867
16	S	Sulfur	[Ne] $3s^2\ 3p^4$	$^3P_2$	10.3600
17	Cl	Chlorine	[Ne] $3s^2\ 3p^5$	$^2P^o_{3/2}$	12.9676
18	Ar	Argon	[Ne] $3s^2\ 3p^6$	$^1S_0$	15.7596
19	K	Potassium	[Ar] $4s$	$^2S_{1/2}$	4.3407
20	Ca	Calcium	[Ar] $4s^2$	$^1S_0$	6.1132
21	Sc	Scandium	[Ar] $3d\ 4s^2$	$^2D_{3/2}$	6.5615
22	Ti	Titanium	[Ar] $3d^2\ 4s^2$	$^3F_2$	6.8281
23	V	Vanadium	[Ar] $3d^3\ 4s^2$	$^4F_{3/2}$	6.7462
24	Cr	Chromium	[Ar] $3d^5\ 4s$	$^7S_3$	6.7665
25	Mn	Manganese	[Ar] $3d^5\ 4s^2$	$^6S_{5/2}$	7.4340
26	Fe	Iron	[Ar] $3d^6\ 4s^2$	$^5D_4$	7.9024
27	Co	Cobalt	[Ar] $3d^7\ 4s^2$	$^4F_{9/2}$	7.8810
28	Ni	Nickel	[Ar] $3d^8\ 4s^2$	$^3F_4$	7.6398
29	Cu	Copper	[Ar] $3d^{10}\ 4s$	$^2S_{1/2}$	7.7264
30	Zn	Zinc	[Ar] $3d^{10}\ 4s^2$	$^1S_0$	9.3942
31	Ga	Gallium	[Ar] $3d^{10}\ 4s^2\ 4p$	$^2P^o_{1/2}$	5.9993
32	Ge	Germanium	[Ar] $3d^{10}\ 4s^2\ 4p^2$	$^3P_0$	7.8994
33	As	Arsenic	[Ar] $3d^{10}\ 4s^2\ 4p^3$	$^4S^o_{3/2}$	9.7886
34	Se	Selenium	[Ar] $3d^{10}\ 4s^2\ 4p^4$	$^3P_2$	9.7524
35	Br	Bromine	[Ar] $3d^{10}\ 4s^2\ 4p^5$	$^2P^o_{3/2}$	11.8138
36	Kr	Krypton	[Ar] $3d^{10}\ 4s^2\ 4p^6$	$^1S_0$	13.9996
37	Rb	Rubidium	[Kr] $5s$	$^2S_{1/2}$	4.1771
38	Sr	Strontium	[Kr] $5s^2$	$^1S_0$	5.6949
39	Y	Yttrium	[Kr] $4d\ 5s^2$	$^2D_{3/2}$	6.2171
40	Zr	Zirconium	[Kr] $4d^2\ 5s^2$	$^3F_2$	6.6339
41	Nb	Niobium	[Kr] $4d^4\ 5s$	$^6D_{1/2}$	6.7589
42	Mo	Molybdenum	[Kr] $4d^5\ 5s$	$^7S_3$	7.0924
43	Tc	Technetium	[Kr] $4d^5\ 5s^2$	$^6S_{5/2}$	7.28
44	Ru	Ruthenium	[Kr] $4d^7\ 5s$	$^5F_5$	7.3605
45	Rh	Rhodium	[Kr] $4d^8\ 5s$	$^4F_{9/2}$	7.4589
46	Pd	Palladium	[Kr] $4d^{10}$	$^1S_0$	8.3369
47	Ag	Silver	[Kr] $4d^{10}\ 5s$	$^2S_{1/2}$	7.5762
48	Cd	Cadmium	[Kr] $4d^{10}\ 5s^2$	$^1S_0$	8.9938
49	In	Indium	[Kr] $4d^{10}\ 5s^2\ 5p$	$^2P^o_{1/2}$	5.7864
50	Sn	Tin	[Kr] $4d^{10}\ 5s^2\ 5p^2$	$^3P_0$	7.3439
51	Sb	Antimony	[Kr] $4d^{10}\ 5s^2\ 5p^3$	$^4S^o_{3/2}$	8.6084
52	Te	Tellurium	[Kr] $4d^{10}\ 5s^2\ 5p^4$	$^3p_2$	9.0096
53	I	Iodine	[Kr] $4d^{10}\ 5s^2\ 5p^5$	$^2P^o_{3/2}$	10.4513
54	Xe	Xenon	[Kr] $4d^{10}\ 5s^2\ 5p^6$	$^1S_0$	12.1298
55	Cs	Cesium	[Xe] $6s$	$^2S_{1/2}$	3.8939
56	Ba	Barium	[Xe] $6s^2$	$^1S_0$	5.2117
57	La	Lanthanum	[Xe] $5d\ 6s^2$	$^2D_{3/2}$	5.5769
58	Ce	Cerium	[Xe] $4f\ 5d\ 6s^2$	$^1G^o_4$	5.5387

Continues

TABLE III (*Continued*)

Z		Element	Ground-state configuration	Ground level	Ionization energy (eV)
59	Pr	Praseodymium	[Xe] $4f^3\ 6s^2$	$^4I^o_{9/2}$	5.473
60	Nd	Neodymium	[Xe] $4f^4\ 6s^2$	$^5I_4$	5.5250
61	Pm	Promethium	[Xe] $4f^5\ 6s^2$	$^6H^o_{5/2}$	5.582
62	Sm	Samarium	[Xe] $4f^6\ 6s^2$	$^7F_0$	5.6436
63	Eu	Europium	[Xe] $4f^7\ 6s^2$	$^8S^o_{7/2}$	5.6704
64	Gd	Gadolinium	[Xe] $4f^7\ 5d\ 6s^2$	$^9D^o_2$	6.1501
65	Tb	Terbium	[Xe] $4f^9\ 6s^2$	$^6H^o_{15/2}$	5.8638
66	Dy	Dysprosium	[Xe] $4f^{10}\ 6s^2$	$^5I_8$	5.9389
67	Ho	Holmium	[Xe] $4f^{11}\ 6s^2$	$^4I^o_{15/2}$	6.0215
68	Er	Erbium	[Xe] $4f^{12}\ 6s^2$	$^3H_6$	6.1077
69	Tm	Thulium	[Xe] $4f^{13}\ 6s^2$	$^2F^o_{7/2}$	6.1843
70	Yb	Ytterbium	[Xe] $4f^{14}\ 6s^2$	$^1S_0$	6.2542
71	Lu	Lutetium	[Xe] $4f^{14}\ 5d\ 6s^2$	$^2D_{3/2}$	5.4259
72	Hf	Hafnium	[Xe] $4f^{14}\ 5d^2\ 6s^2$	$^3F_2$	6.8251
73	Ta	Tantalum	[Xe] $4f^{14}\ 5d^3\ 6s^2$	$^4F_{3/2}$	7.5496
74	W	Tungsten	[Xe] $4f^{14}\ 5d^4\ 6s^2$	$^5D_0$	7.8640
75	Re	Rhenium	[Xe] $4f^{14}\ 5d^5\ 6s^2$	$^6S_{5/2}$	7.8335
76	Os	Osmium	[Xe] $4f^{14}\ 5d^6\ 6s^2$	$^5D_4$	8.4382
77	Ir	Iridium	[Xe] $4f^{14}\ 5d^7\ 6s^2$	$^4F_{9/2}$	8.9670
78	Pt	Platinum	[Xe] $4f^{14}\ 5d^9\ 6s$	$^3D_3$	8.9587
79	Au	Gold	[Xe] $4f^{14}\ 5d^{10}\ 6s$	$^2S_{1/2}$	9.2255
80	Hg	Mercury	[Xe] $4f^{14}\ 5d^{10}\ 6s^2$	$^1S_0$	10.4375
81	Tl	Thallium	[Xe] $4f^{14}\ 5d^{10}\ 6s^2\ 6p$	$^2P^o_{1/2}$	6.1082
82	Pb	Lead	[Xe] $4f^{14}\ 5d^{10}\ 6s^2\ 6p^2$	$^3P_0$	7.4167
83	Bi	Bismuth	[Xe] $4f^{14}\ 5d^{10}\ 6s^2\ 6p^3$	$^4S^o_{3/2}$	7.2856
84	Po	Polonium	[Xe] $4f^{14}\ 5d^{10}\ 6s^2\ 6p^4$	$^3P_2$	8.417?
85	At	Astatine	[Xe] $4f^{14}\ 5d^{10}\ 6s^2\ 6p^5$	$^2P^o_{3/2}$	
86	Rn	Radon	[Xe] $4f^{14}\ 5d^{10}\ 6s^2\ 6p^6$	$^1S_0$	10.7485
87	Fr	Francium	[Rn] $7s$	$^2S_{1/2}$	4.0727
88	Ra	Radium	[Rn] $7s^2$	$^1S_0$	5.2784
89	Ac	Actinium	[Rn] $6d\ 7s^2$	$^2D_{3/2}$	5.17
90	Th	Thorium	[Rn] $6d^2\ 7s^2$	$3F_2$	6.3067
91	Pa	Protactinium	[Rn] $5f^2(^3H_4)\ 6d\ 7s^2$	$(4,\frac{3}{2})_{11/2}$	5.89
92	U	Uranium	[Rn] $5f^3(^4I^o_{9/2})\ 6d\ 7s^2$	$(\frac{9}{2},\frac{3}{2})^o_6$	6.1941
93	Np	Neptunium	[Rn] $5f^4(^5I_4)\ 6d\ 7s^2$	$(4,\frac{3}{2})_{11/2}$	6.2657
94	Pu	Plutonium	[Rn] $5f^6\ 7s^2$	$^7F_0$	6.0262
95	Am	Americium	[Rn] $5f^7\ 7s^2$	$^8S^o_{7/2}$	5.9738
96	Cm	Curium	[Rn] $5f^7\ 6d\ 7s^2$	$^9D^o_2$	5.9915
97	Bk	Berkelium	[Rn] $5f^9\ 7s^2$	$^6H^o_{15/2}$	6.1979
98	Cf	Californium	[Rn] $5f^{10}\ 7s^2$	$^5I_8$	6.2817
99	Es	Einsteinium	[Rn] $5f^{11}\ 7s^2$	$^4I^o_{15/2}$	6.42
100	Fm	Fermium	[Rn] $5f^{12}\ 7s^2$	$^3H_6$	6.50
101	Md	Mendelevium	[Rn] $5f^{13}\ 7s^2$	$^2F^o_{7/2}$	6.58
102	No	Nobelium	[Rn] $5f^{14}\ 7s^2$	$^1S_0$	6.65
103	Lr	Lawrencium	[Rn] $5f^{14}\ 7s^2\ 7p$?	$^2P^o_{1/2}$?	4.9?
104	Rf	Rutherfordium	[Rn] $5f^{14}\ 6d^2\ 7s^2$?	$^3F_2$?	6.0?

Source: Martin, W. C., and Musgrove, A. (2001). NIST Physics Reference Data Web Site, www.physics.nist.gov/PhysRefData/.

laboratory from which the data came is also a factor, since some research groups are known to take greater care in their work than others.

When there is a high degree of interrelation among a set of independently measured quantities, a systematic correlation scheme can be devised. Thermodynamics provides the prime example. Here one may have available calorimetric measurements of enthalpy changes in chemical reactions, heat capacity measurements, equilibrium constants as a function of temperature, entropy calculated from molecular constants, and perhaps other pertinent experimental measurements. When reduced to standard temperature and pressure, all the data relevant to a given reaction must satisfy well-established thermodynamic

TABLE IV Properties of Selected Nuclides

	Abundance or half-life	Atomic mass (u)	Mass excess (keV)	Spin	Magnetic moment (μ_N)	Quadrupole moment (fm^2)
^{1}n	10.3 m	1.008 664 916	8071.317	1/2	−1.91304272	
^{1}H	99.985%	1.007 825 032	7288.969	1/2	+2.7928473	
^{2}H	0.015%	2.014 101 778	13135.720	1	+0.8574382	+0.286
^{3}H	12.32 y	3.016 049 268	14949.794	1/2	+2.9789625	
^{3}He	0.000137%	3.016 029 310	14931.204	1/2	−2.1276248	
^{4}He	99.999863%	4.002 603 250	2424.911	0	0	
^{6}Li	7.5%	6.015 122 3	14086.312	1	+0.8220467	−0.082
^{7}Li	92.5%	7.016 004 0	14907.673	3/2	+3.256427	−4.01
^{9}Be	100%	9.012 182 1	11347.584	3/2	−1.1779	+5.288
^{10}B	19.9%	10.012 937 0	12050.761	3	+1.800645	+8.459
^{11}B	80.1%	11.009 305 5	8667.984	3/2	+2.688649	+4.059
^{12}C	98.90%	12	0	0	0	
^{13}C	1.10%	13.003 354 838	3125.011	1/2	+0.7024118	
^{14}C	5715 y	14.003 241 988	3019.892	0	0	
^{14}N	99.634%	14.003 074 005	2863.417	1	+0.4037610	+2.02
^{15}N	0.366%	15.000 108 898	101.438	1/2	−0.2831888	
^{16}O	99.762%	15.994 914 622	−4736.998	0	0	
^{19}F	100%	18.998 403 21	−1487.405	1/2	+2.628868	
^{23}Na	100%	22.989 769 7	−9529.485	3/2	+2.217522	+10.89
^{31}P	100%	30.973 761 5	−24440.991	1/2	+1.13160	
^{32}S	95.02%	31.972 070 7	−26015.981	0	0	
^{34}S	4.21%	33.967 866 8	−29931.850	0	0	
^{55}Fe	2.73 y	54.938 298 029	−57475.007	3/2		
^{60}Co	5.271 y	59.933 822 196	−61644.218	5	+3.799	+44
^{90}Sr	29.1 y	89.907 737 596	−85941.863	0		
^{131}I	8.040 d	130.906 124 168	−87444.761	7/2	+2.742	−40
^{222}Rn	3.8235 d	222.017 570	16366.787	0	0	
^{226}Ra	1599 y	226.025 403	23662.324	0	0	
^{235}U	0.7200%	235.043 923	40914.062	7/2	−0.38	+493.6
^{238}U	99.2745%	238.050 783	47303.664	0	0	
^{239}Pu	24110 y	239.052 157	48583.478	1/2	+0.203	

Source: Lide, D. R., ed. (1999). "CRC Handbook of Chemistry and Physics," CRC Press, Boca Raton, FL.

relations. Furthermore, the energy and entropy changes for a process must be independent of the path followed. These constraints enable one to check the internal consistency of large data sets whose individual values come from a variety of sources. In this way, faulty measurements are frequently recognized that would not be suspected if examined in isolation.

Chemical thermodynamic data and thermophysical properties of fluids are routinely evaluated in this manner. Computer programs have been developed to assess large data sets and select recommended values through a least-squares or similar fitting procedure. Other fields amenable to this approach are atomic and molecular spectroscopy, nuclear physics, and crystallography. In still other cases, such as chemical kinetics and atomic collision cross sections, theory can be used to place limits on data values.

Ideally, the aim of every evaluation effort is to present a "best" or "recommended" value plus a quantitative statement of its uncertainty. If the dominant errors are truly random, a standard deviation or 95% confidence interval can be quoted, which gives the user a sound basis for deciding the implication of this uncertainty for a given problem. However, this situation almost never applies; instead, the most significant errors are usually systematic in nature, deriving from either the initial measurement process or the model used in analyzing the data. The correlations of large data sets described above are very helpful in uncovering such systematic errors, but the judgment of an experienced researcher is also extremely important.

TABLE V Specific Heat, Thermal Conductivity, and Coefficient of Thermal Expansion of the Solid Elements at 25°C

Element	c_p (J g^{-1} K^{-1})	λ (W cm^{-1} K^{-1})	α (10^{-6} K^{-1})
Aluminum	0.897	2.37	23.1
Antimony	0.207	0.24	11.0
Arsenic	0.329	0.50	15.5
Barium	0.204	0.18	20.6
Beryllium	1.825	2.00	11.3
Bismuth	0.122	0.08	13.4
Boron	1.026	0.27	4.7
Cadmium	0.232	0.97	30.8
Calcium	0.647	2.00	22.3
Carbon (diamond)	0.509	9.00	1.1
Cerium	0.192	0.11	5.2
Cesium	0.242	0.36	—
Chromium	0.449	0.94	4.9
Cobalt	0.421	1.00	13.0
Copper	0.385	4.01	16.5
Dysprosium	0.173	0.11	9.9
Erbium	0.168	0.15	12.2
Europium	0.182	0.14	35.0
Gadolinium	0.236	0.11	9.4
Gallium	0.371	0.41	—
Germanium	0.320	0.60	5.8
Gold	0.129	3.17	14.2
Hafnium	0.144	0.23	5.9
Holmium	0.165	0.16	11.2
Indium	0.233	0.82	32.1
Iridium	0.131	1.47	6.4
Iron	0.449	0.8	11.8
Lanthanum	0.195	0.13	12.1
Lead	0.129	0.35	28.9
Lithium	3.582	0.85	46
Lutetium	0.154	0.16	9.9
Magnesium	1.023	1.56	24.8
Manganese	0.479	0.08	21.7
Mercury	0.140	0.08	—
Molybdenum	0.251	1.38	4.8
Neodymium	0.190	0.17	9.6
Nickel	0.444	0.91	13.4
Niobium	0.265	0.54	7.3
Osmium	0.130	0.88	5.1
Palladium	0.246	0.72	11.8
Phosphorus (white)	0.769	0.24	—
Platinum	0.133	0.72	8.8
Plutonium	—	0.07	46.7
Potassium	0.757	1.02	—
Praseodymium	0.193	0.13	6.7
Promethium	—	0.15	11
Rhenium	0.137	0.48	6.2
Rhodium	0.243	1.50	8.2
Rubidium	0.363	0.58	—
Ruthenium	0.238	1.17	6.4
Samarium	0.197	0.13	12.7
Scandium	0.568	0.16	10.2
Silicon	0.705	1.48	2.6
Silver	0.235	4.29	18.9
Sodium	1.228	1.41	71
Strontium	0.301	0.35	22.5
Sulfur (rhombic)	0.710	0.27	—
Tantalum	0.140	0.58	6.3
Technetium	—	0.51	—
Terbium	0.182	0.11	10.3
Thallium	0.129	0.46	29.9
Thorium	0.113	0.540	11.0
Thulium	0.160	0.17	13.3
Tin	0.228	0.67	22.0
Titanium	0.523	0.22	8.6
Tungsten	0.132	1.74	4.5
Uranium	0.116	0.28	13.9
Vanadium	0.489	0.31	8.4
Ytterbium	0.155	0.39	26.3
Yttrium	0.298	0.17	10.6
Zinc	0.388	1.16	30.2
Zirconium	0.278	0.227	5.7

Source: Adapted from Anderson, H. L., ed. (1989). "A Physicist's Desk Reference," Springer-Verlag, New York; with updates from Lide, D. R., ed. (1999). "CRC Handbook of Chemistry and Physics," CRC Press, Boca Raton, FL.

V. DISSEMINATION OF CRITICAL DATA

Traditionally, books and journals have served as the major vehicles for disseminating critically evaluated data to scientists and engineers. Several widely used series of tables have already been mentioned. The *Journal of Physical and Chemical Reference Data*, published jointly by the American Institute of Physics and the National Institute of Standards and Technology, is one of the major vehicles for disseminating tables of recommended data and documenting the methodology used for their evaluation. This journal is published bimonthly, with supplements appearing on an irregular basis. More specialized data journals also exist—for example, *Atomic Data and Nuclear Data Tables* (Academic Press) and *Journal of Phase Equilibria* (ASM International). Finally, many technical publishers offer monographs and handbooks containing evaluated data on specialized subjects.

TABLE VI Vapor Pressure of the Elements

		Temperature (°C) for the indicated pressure[a]					
	Element	**1 Pa**	**10 Pa**	**100 Pa**	**1 kPa**	**10 kPa**	**100 kPa**
Ag	Silver	1010	1140	1302	1509	1782	2160
Al	Aluminum	1209	1359	1544	1781	2091	2517
Ar	Argon	—	−226.4 s	−220.3 s	−212.4 s	−201.7 s	−186.0
As	Arsenic	280 s	323 s	373 s	433 s	508 s	601 s
At	Astatine	88 s	119 s	156 s	202 s	258 s	334
Au	Gold	1373	1541	1748	2008	2347	2805
B	Boron	2075	2289	2549	2868	3272	3799
Ba	Barium	638 s	765	912	1115	1413	1897
Be	Beryllium	1189 s	1335	1518	1750	2054	2469
Bi	Bismuth	668	768	892	1052	1265	1562
Br_2	Bromine	−87.7 s	−71.8 s	−52.7 s	−29.3 s	2.5	58.4
C	Carbon (graphite)	—	2566 s	2775 s	3016 s	3299 s	3635 s
Ca	Calcium	591 s	683 s	798 s	954	1170	1482
Cd	Cadmium	257 s	310 s	381	472	594	767
Ce	Cerium	1719	1921	2169	2481	2886	3432
Cl_2	Chlorine	−145 s	−133.7 s	−120.2 s	−103.6 s	−76.1	−34.2
Co	Cobalt	1517	1687	1892	2150	2482	2925
Cr	Chromium	1383 s	1534 s	1718 s	1950	2257	2669
Cs	Cesium	144.5	195.6	260.9	350.0	477.1	667.0
Cu	Copper	1236	1388	1577	1816	2131	2563
Dy	Dysprosium	1105 s	1250 s	1431	1681	2031	2558
Er	Erbium	1231 s	1390 s	1612	1890	2279	2859
Eu	Europium	590 s	684 s	799 s	961	1179	1523
F_2	Fluorine	−235 s	−229.5 s	−222.9 s	−214.8	−204.3	−188.3
Fe	Iron	1455 s	1617	1818	2073	2406	2859
Fr	Francium	131	181	246	335	465	673
Ga	Gallium	1037	1175	1347	1565	1852	2245
Gd	Gadolinium	1563	1755	1994	2300	2703	3262
Ge	Germanium	1371	1541	1750	2014	2360	2831
H_2	Hydrogen	—	—	—	—	−258.6	−252.8
He	Helium	—	—	—	—	−270.6	−268.9
Hf	Hafnium	2416	2681	3004	3406	3921	4603
Hg	Mercury	42.0	76.6	120.0	175.6	250.3	355.9
Ho	Holmium	1159 s	1311 s	1502	1767	2137	2691
I_2	Iodine	−12.8 s	9.3 s	35.9 s	68.7 s	108 s	184.0
In	Indium	923	1052	1212	1417	1689	2067
Ir	Iridium	2440 s	2684	2979	3341	3796	4386
K	Potassium	200.2	256.5	328	424	559	756.2
Kr	Krypton	−214.0 s	−208.0 s	−199.4 s	−188.9 s	−174.6 s	−153.6
La	Lanthanum	1732	1935	2185	2499	2905	3453
Li	Lithium	524.3	612.3	722.1	871.2	1064.3	1337.1
Lu	Lutetium	1633 s	1829.8	2072.8	2380	2799	3390
Mg	Magnesium	428 s	500 s	588 s	698	859	1088
Mn	Manganese	955 s	1074 s	1220 s	1418	1682	2060
Mo	Molybdenum	2469 s	2721	3039	3434	3939	4606
N_2	Nitrogen	−236 s	−232 s	−226.8 s	−220.2 s	−211.1 s	−195.9
Na	Sodium	280.6	344.2	424.3	529	673	880.2

Continues

TABLE VI (*Continued*)

		Temperature (°C) for the indicated pressure[a]					
	Element	**1 Pa**	**10 Pa**	**100 Pa**	**1 kPa**	**10 kPa**	**100 kPa**
Nb	Niobium	2669	2934	3251	3637	4120	4740
Nd	Neodymium	1322.3	1501.2	1725.3	2023	2442	3063
Ne	Neon	−261 s	−260 s	−258 s	−255 s	−252 s	−246.1
Ni	Nickel	1510	1677	1881	2137	2468	2911
O_2	Oxygen	—	—	—	−211.9	−200.5	−183.1
Os	Osmium	2887 s	3150	3478	3875	4365	4983
P	Phosphorus (white)	6 s	34 s	69	115	180	276
P	Phosphorus (red)	182 s	216 s	256 s	303 s	362 s	431 s
Pb	Lead	705	815	956	1139	1387	1754
Pd	Palladium	1448 s	1624	1844	2122	2480	2961
Po	Polonium	—	—	—	573	730.2	963.3
Pr	Praseodymium	1497.7	1699.4	1954	2298	2781	3506
Pt	Platinum	2057	2277	2542	2870	3283	3821
Pu	Plutonium	1483	1680	1925	2238	2653	3226
Ra	Radium	546 s	633 s	764	936	1173	1526
Rb	Rubidium	160.4	212.5	278.9	368	496.1	685.3
Re	Rhenium	3030 s	3341	3736	4227	4854	5681
Rh	Rhodium	2015	2223	2476	2790	3132	3724
Rn	Radon	−163 s	−152 s	−139 s	−121.4 s	−97.6 s	−62.3
Ru	Ruthenium	2315 s	2538	2814	3151	3572	4115
S	Sulfur	102 s	135	176	235	318	444
Sb	Antimony	534 s	603 s	738	946	1218	1585
Sc	Scandium	1372 s	1531 s	1733	1993	2340	2828
Se	Selenium	227	279	344	431	540	685
Si	Silicon	1635	1829	2066	2363	2748	3264
Sm	Samarium	728 s	833 s	967 s	1148	1402	1788
Sn	Tin	1224	1384	1582	1834	2165	2620
Sr	Strontium	523 s	609 s	717 s	866	1072	1373
Ta	Tantalum	3024	3324	3684	4122	4666	5361
Tb	Terbium	1516.1	1706.1	1928	2232	2640	3218
Tc	Technetium	2454	2725	3051	3453	3961	4621
Te	Tellurium	—	—	502	615	768.8	992.4
Th	Thorium	2360	2634	2975	3410	3986	4782
Ti	Titanium	1709	1898	2130	2419	2791	3285
Tl	Thallium	609	704	824	979	1188	1485
Tm	Thulium	844 s	962 s	1108 s	1297 s	1548	1944
U	Uranium	2052	2291	2586	2961	3454	4129
V	Vanadium	1828 s	2016	2250	2541	2914	3406
W	Tungsten	3204 s	3500	3864	4306	4854	5550
Xe	Xenon	−190 s	−181 s	−170 s	−155.8 s	−136.6 s	−108.4
Y	Yttrium	1610.1	1802.3	2047	2354	2763	3334
Yb	Ytterbium	463 s	540 s	637 s	774 s	993	1192
Zn	Zinc	337 s	397 s	477	579	717	912
Zr	Zirconium	2366	2618	2924	3302	3780	4405

[a] An "s" following an entry indicates the substance is solid at that temperature.

Source: Lide, D. R., ed. (1999). "CRC Handbook of Chemistry and Physics," CRC Press, Boca Raton, FL.

TABLE VII Properties of Some Common Fluids

Formula	Fluid	Normal melting point (t_m/°C)	Normal boiling point (t_b/°C)	Critical constants (t_c/°C)	(p_c/°C)
He	Helium	—	–268.93	–267.96	0.23
Ar	Argon	–189.36[a]	–185.85	–122.28	4.9
H_2	Hydrogen	–259.34	–252.87	–240.18	1.29
O_2	Oxygen	–218.79	–182.95	–118.56	5.04
N_2	Nitrogen	–210	–195.79	–146.94	3.39
CO	Carbon monoxide	–205.02	–191.5	–140.24	3.5
CO_2	Carbon dioxide	–56.56[a]	–78.4[b]	30.98	7.38
H_2O	Water	0.00	100.0	373.99	22.06
NH_3	Ammonia	–77.73	–33.33	132.4	11.35
N_2O	Nitrous oxide	–90.8	–88.48	36.42	7.26
CH_4	Methane	–182.47	–161.4	–82.59	4.6
C_2H_6	Ethane	–182.79	–88.6	32.17	4.87
C_3H_8	Propane	–187.63	–42.1	96.68	4.25
C_4H_{10}	Butane	–138.3	–0.5	151.97	3.8
C_2H_4	Ethylene	–169.15	–103.7	9.19	5.04
C_6H_6	Benzene	5.49	80.09	288.9	4.9
CH_4O	Methanol	–97.53	64.6	239.4	8.08
C_2H_6O	Ethanol	–114.14	78.29	240.9	6.14
C_3H_6O	Acetone	–94.7	56.05	235.0	4.700

[a] Solid–liquid–gas triple point.
[b] Sublimation point, where vapor pressure of solid reaches 1 atm.
Source: Lide, D. R., ed. (1999). "CRC Handbook of Chemistry and Physics," CRC Press, Boca Raton, FL.

TABLE VIII CODATA Key Values for Thermodynamics

Substance	State	Relative molecular mass	$\Delta_f H°$(298.15 K) (kJ mol^{-1})	$S°$(298.15 K) (J K^{-1} mol^{-1})	$H°$(298.15 K) − $H°$(0) (kJ mol^{-1})
O	Gas	15.9994	249.18 ± 0.10	160.950 ± 0.003	6.725 ± 0.001
O_2	Gas	31.9988	0	205.043 ± 0.005	8.680 ± 0.002
H	Gas	1.00794	217.998 ± 0.006	114.608 ± 0.002	6.197 ± 0.001
H^+	Aqueous	1.0074	0	0	—
H_2	Gas	2.0159	0	130.571 ± 0.005	8.468 ± 0.001
OH^-	Aqueous	17.0079	–230.015 ± 0.040	–10.90 ± 0.20	—
H_2O	Liquid	18.0153	–285.830 ± 0.040	69.95 ± 0.03	13.273 ± 0.020
H_2O	Gas	18.0153	–241.826 ± 0.040	188.726 ± 0.010	9.905 ± 0.005
He	Gas	4.00260	0	126.044 ± 0.002	6.197 ± 0.001
Ne	Gas	20.179	0	146.219 ± 0.003	6.197 ± 0.001
Ar	Gas	39.948	0	154.737 ± 0.003	6.197 ± 0.001
Kr	Gas	83.80	0	163.976 ± 0.003	6.197 ± 0.001
Xe	Gas	131.29	0	169.576 ± 0.003	6.197 ± 0.001
F	Gas	18.99840	79.38 ± 0.30	158.642 ± 0.004	6.518 ± 0.001
F^-	Aqueous	18.9989	–335.35 ± 0.65	–13.8 ± 0.8	—
F_2	Gas	37.9968	0	202.682 ± 0.005	8.825 ± 0.001
HF	Gas	20.0063	−273.30 ± 0.70	173.670 ± 0.003	8.599 ± 0.001
Cl	Gas	35.453	121.301 ± 0.008	165.081 ± 0.004	6.272 ± 0.001
Cl^-	Aqueous	35.4535	–167.080 ± 0.10	56.60 ± 0.20	—

Source: Excerpted from Cox, J. D., Wagman, D. D., and Medvedev, V. A. (1989). "CODATA Key Values for Thermodynamics," Hemisphere Publishing, New York.

There are also many handbooks with a broad coverage of physical and chemical data; among the most familiar of these are the *CRC Handbook of Chemistry and Physics*, *The Merck Index*, and the *American Institute of Physics Handbook*. Such handbooks are very convenient data sources. While they cannot provide the backup documentation found in the data journals and monographs discussed above, the better ones carry references to more detailed publications.

The decade beginning in 1990 saw a major change in the manner of disseminating all types of information, and scientific data were no exception. There are many advantages associated with computerized data dissemination. One consideration is economic. While the costs incurred with composition and printing have continued to increase, computer costs for data storage and network communications have decreased sharply, thus making the electronic dissemination of critical data more attractive. Often the sheer volume of data makes a machine-readable format the only practical way of storage and distribution. Electronic databases lend themselves to easy updating, thus promoting the currency of the data, and search and retrieval are far more powerful. Having the data in electronic form also makes it easier for the user to carry out calculations and look for trends that may lead to new scientific insights.

Although these advantages were recognized much earlier, the transition to electronic dissemination of scientific data did not begin to accelerate until the mid-1990s. Two factors have contributed: the expanding availability of personal computers with CD ROM drives and the explosive growth of the Internet. The CD ROM has proved to be an efficient means for distributing physical and chemical databases and the accompanying software to individuals for use on their personal computers. The Internet provides an inexpensive way for users to access large databases maintained on institutional computers. The graphical capabilities of the World Wide Web have also contributed by making it easy to display special characters, chemical structures, and other non-text information. Finally, the growing use of computers for data analysis, process simulation, engineering design, and similar applications has created a demand for data in digital, as opposed to paper, format.

The ease with which information can be posted on the Internet has had one unfortunate consequence. There are a great many sites that purport to provide physical and chemical data, but the quality is highly variable. Data quality is a consideration even when dealing with printed compilations, but on the Internet the traditional filter of the publication process can no longer be relied upon. A search for a specific property on a standard Internet search engine is likely to turn up hundreds of sites, most of which have no documentation and provide no basis for confidence in the correctness of the data presented. It is important for all users of data taken from the Internet to evaluate the reliability of the source and assure that it is truly critical data.

Some of the important World Wide Web sites for evaluated physical and chemical data are listed below:

- NIST Physics Data, covering fundamental constants, atomic spectra, and X-ray data; <physics.nist.gov>
- Fundamental Particle Properties, prepared by the Particle Data Group at Lawrence Berkeley Laboratories; <pdg.lbl.gov>
- NIST Chemistry Webbook, whose topics include thermodynamics, ion energetics, infrared and mass spectra, fluid properties, etc; <webbook.nist.gov>
- Beilstein and Gmelin Databases, covering chemical properties of organic and inorganic compounds; <www.beilstein.com>
- Hazardous Substances Data Bank, maintained by the National Library of Medicine and containing physical property data as well as toxicity and safety data; <chem.sis.nlm.nih.gov/hsdb/>
- CRCnetBase, including the Web version of the CRC Handbook of Chemistry and Physics, The Merck Index, and other databases; <www.crcpress.com>

Crystallographic databases are maintained for different classes of materials:

- Organic compounds: <www.ccdc.cam.ac.uk>
- Inorganic compounds: <www.nist.gov/srd/> and <www.fiz-karlsruhe.de>
- Metals: <www.tothcanada.com>
- Proteins: <www.rcsb.org>
- Nucleic acids: <www.ndbserver.rutgers.edu>

SEE ALSO THE FOLLOWING ARTICLES

CHEMICAL THERMODYNAMICS • DATABASES • MECHANICS, CLASSICAL • PERIODIC TABLE (CHEMISTRY) • THERMAL ANALYSIS • THERMOMETRY

BIBLIOGRAPHY

Anderson, H. L., ed. (1989). "A Physicist's Desk Reference," 2nd ed., Amer. Inst. of Phys., New York.

Dubois, J.-E., and Gershon, N., eds. (1996). "The Information Revolution: Impact on Science and Technology," Springer-Verlag, Berlin.

Glaeser, P. S., ed. (1992). "Data for Discovery," Begell House, New York.

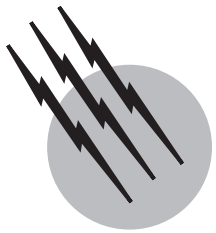
Lide, D. R. (1973). "Status report on critical compilation of physical chemical data." *Ann. Rev. Phys. Chem.* **24,** 135–158.

Lide, D. R. (1981). "Critical data for critical needs." *Science* **212,** 135–158.

Maizell, R. E. (1998). "How To Find Chemical Information," Wiley-Interscience, New York.

Molino, B. B. (1985). *In* "The Role of Data in Scientific Progress" (P. S. Glaser, ed.), North-Holland, Amsterdam.

Rumble, J. R., and Hampel, V. E. (1984). "Database Management in Science and Technology," North-Holland, Amsterdam.



Cryogenics

P. V. E. McClintock
University of Lancaster

GLOSSARY

Boson Entity whose intrinsic spin is an even multiple of $\hbar$, examples being photons, phonons, and atoms made up of an even number of fundamental particles.

Bose–Einstein statistics Form of quantum statistics that must be used to describe a gaseous assembly of noninteracting bosons at low temperatures.

Cryostat Apparatus used to achieve and maintain cryogenic temperatures.

Debye cut-off frequency Maximum possible frequency for a vibrational wave in a crystal: the Debye characteristic temperature is the Debye cut-off frequency multiplied by Planck's constant and divided by Boltzmann's constant.

Dewar Vacuum-insulated vessel, of a type commonly used for cryostats or as containers for liquefied gases.

Dispersion curve Frequencies (energies) of the permitted excitations of a system plotted as a function of their momenta (wave vectors).

Fermi–Dirac statistics Form of quantum statistics that must be used to describe a gaseous assembly of noninteracting fermions at low temperatures.

Fermi sphere Sphere of filled states in momentum space, characteristic of an ideal Fermi–Dirac gas at a temperature very much less than the Fermi temperature.

Fermi surface Surface of the Fermi sphere, which is a region of particular importance since it relates to the only particles that can readily undergo interactions; particles on the Fermi surface at very low temperatures have the Fermi momentum and Fermi energy; the Fermi temperature is the Fermi energy divided by Boltzmann's constant.

Fermion Entity whose intrinsic spin is a halfintegral multiple of $\hbar$, examples being electrons, protons, neutrons, and atoms made up of an odd number of fundamental particles.

He I Nonsuperfluid, higher temperature phase of liquid ^{4}He.

He II Superfluid, lower temperature phase of liquid ^{4}He.

^{3}He-A, ^{3}He-B Superfluid phases of liquid ^{3}He.

Phonon Quantum of vibrational energy in a crystal or in He II; phonons can be envisaged as quasi-particles traveling at the velocity of sound.

Roton Excitation, somehow analogous to, but different from, a phonon, near the minimum of the dispersion curve of He II.

Superconductivity State of zero electrical resistivity and perfect diamagnetism in a material.

Superfluidity State of a fluid in which it has zero viscosity, examples being He II, ^{3}He-A, ^{3}He-B, and the electron gas in a superconductor.

Superleak Material, usually a tightly packed powder or a medium with very fine pores, through which a superfluid can pass easily but which is impermeable to any normal (viscous) fluid.

CRYOGENICS is the science and technology of very low temperatures. Although there is no hard and fast definition of where ordinary refrigeration gives way to cryogenics, it is customary to consider the dividing line as being at ~100 K. Of particular scientific interest and importance are the cryogenic phenomena of superconductivity (below ~120 K) and superfluidity (below ~2 K), for which no known analogues exist in the everyday world at room temperature.

Cryogenic techniques are now being employed in applications as diverse as medicine, rocketry, archeology, and metrology.

I. NATURE AND SIGNIFICANCE OF LOW TEMPERATURES

A. Temperature

It is now possible to study the properties of matter experimentally over some 15 decades in temperature. That is, the highest temperature attainable is about 10^{15} times larger than the lowest one. As indicated in Fig. 1, it is feasible to reach temperatures as low as 10^{-7} K (for copper nuclei) with state-of-the-art cryogenic technology, or as high as 10^8 K in the case of hydrogen fusion experiments in a tokamak, starting in each case from the ambient temperature of 3×10^2 K. Important scientific insights into the fundamental nature of matter, with corresponding innovations and applications in technology, have arisen from continued progress at both ends of the temperature scale (and it is interesting to note that low-temperature physicists and engineers have already traveled further from ambient than their high-temperature colleagues by a factor of 10^4). In this article, we concentrate on what happens in the lower temperature two-thirds of Fig. 1.

The Kelvin temperature scale is absolute in the sense that it does not depend on any particular property of any particular material (the ratio of two temperatures being formally defined as the ratio of the heats accepted and rejected by an ideal engine operating between thermal reservoirs at those temperatures). The size of the degree Kelvin (K)

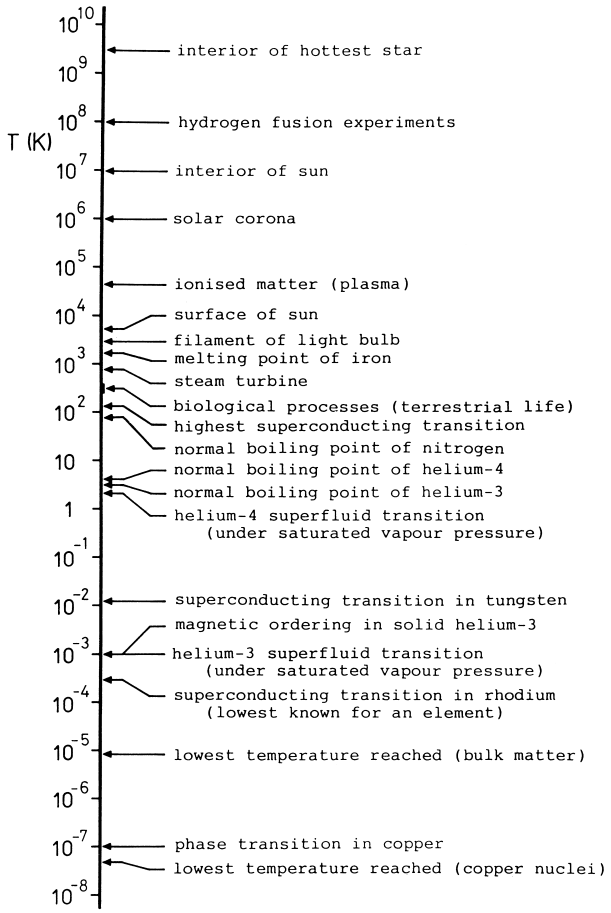


FIGURE 1 The temperatures T (in Kelvin) at which some selected phenomena occur. Note that the scale is logarithmic.

has been chosen, however, so as to be consistent with the degree Celsius of the earlier ideal gas temperature scale.

There is an absolute zero of temperature: 0 K on the Kelvin scale. In terms of classical concepts this would correspond to a complete cessation of all atomic motion. In reality, however, this is far from being the case. Real systems are governed by quantum mechanics for which the ground state does not have zero energy. The quantity that does approach zero as $T \to 0$ is the *entropy*, which gives a measure of the disorder of the system. The third law of thermodynamics, embodying this observation, is of particular importance for cryogenics. It can be stated in two forms. The first of these is

The entropy of all systems and of all states of a system is zero at absolute zero,

which, it should be noted, is subject to certain restrictions concerning the thermodynamic reversibility of connecting paths between the states of the system. The second form of the law is of more immediately obvious relevance:

It is impossible to reach the absolute zero of temperature by any finite number of processes.

In other words, the absolute zero is an unattainable goal to which the cryogenic engineer may aspire but which he or she can, by definition, never quite reach. The logarithmic temperature scale of Fig. 1 provides a convenient reminder of the restrictive nature of the third law (since 0 K would be situated an infinite distance downwards on such a plot). It is also appropriate, in the sense that the number of interesting changes in properties that occur when the temperature is altered often seems to depend more on the *factor* by which T changes than on the absolute magnitude of any change.

B. Matter at Low Temperatures

It is scarcely surprising that the properties of all materials undergo enormous changes as they are cooled through the huge range of low temperatures that is accessible. We discuss some particular cryogenic phenomena below; in this section we will try to provide a general overview of the sorts of changes in properties that occur in practice between, say, room temperature and 1 K.

The first and most striking difference between materials at room temperature and at 1 K is that (with the sole exception of helium) bulk matter at 1 K exists only in solid form. There are no liquids or gases.

The specific heats of most materials at 1 K are very small compared to their values at room temperature. This is, in fact, only to be expected because the specific heat at constant volume may be written in terms of the entropy S as:

$$C_v = T(\partial S/\partial T)_v$$

and the third law tells us that entropy changes all tend to zero as the temperature T tends to zero. Thus, at a sufficiently low temperature, all specific heats must become negligible.

The coefficient of thermal conductivity κ usually varies with T in quite a complicated way (see below) and at 1 K it may be larger or smaller than it is at 300 K, depending on the material. The coefficient of thermal diffusion, given by:

$$k_{\mathrm{D}} = \kappa/C_v\rho$$

where ρ is the density, is usually much larger than at 300 K, however, which causes thermal equilibrium times at 1 K to be very short.

The electrical resistivity of pure metals usually falls by orders of magnitude on their being cooled from 300 K to 1 K. That of alloys, on the other hand, tends to be relatively temperature independent. In both cases, there is a possible exception. At a range of critical temperatures below about 20 K for metals and alloys, or below about 120 K for the high $T_{\hat{c}}$ oxide systems (see Fig. 1), the electrical resistivity of certain materials drops discontinuously to zero; they are then said to have become *superconductors*.

The coefficient of thermal expansion β decreases with T and, in the design of cryostats and other low-temperature apparatus, it can usually be ignored for temperatures below the normal boiling point of nitrogen (77 K). At 1 K, β is negligible. It can be shown that the third law requires that $\beta \to 0$ as $T \to 0$.

The behavior of helium at low temperatures is quite different from that of all other materials. When cooled under atmospheric pressure, helium liquefies at 4.2 K, but it never solidifies, no matter how cold it is made. This behavior is inexplicable in terms of classical physics where, at a sufficiently low temperature, even the very weak interatomic forces that exist between helium atoms should be sufficient to draw the material together into a crystalline solid. Furthermore, the liquid in question has many extraordinary properties. In particular, it undergoes a transition at ~2 K to a state of *superfluidity*, such that it has zero viscosity and can flow without dissipation of energy, even through channels of vanishingly small dimensions. This frictionless flow of the liquid is closely analogous to the frictionless flow of the electrons in a superconductor. On Earth, superfluidity and superconductivity are exclusively low-temperature phenomena, but it is inferred that they probably also arise in the proton and neutron fluids within neutron stars.

We have been referring, thus far, to the common isotope of helium, ^{4}He. The rare isotope. ^{3}He, behaves in an equally extraordinary manner at low temperatures but with some interesting and important differences to which we will return, below. When cooled under atmospheric pressure, ^{3}He liquefies at 3.2 K but it does not undergo a superfluid transition until the temperature has fallen to 1 mK.

Superfluidity and superconductivity represent particularly dramatic manifestations of the quantum theory on a grand scale and, in common also with virtually all other cryogenic phenomena, they can only be understood satisfactorily in terms of quantum statistical mechanics.

C. Quantum Statistics

The behavior of materials at cryogenic temperatures is dominated by quantum effects. It matters, to a very much greater extent than at room temperature, that the energy levels available to a system are in reality discrete and not continuous. The symmetry of the wave functions of the constituent particles of the system is also of crucial importance since it is this that determines whether or not the

occupation of a given quantum state by more than one particle is possible.

Many cryogenic systems can be modeled successfully as gases, even including some that appear at first sight to be solids or liquids, as we shall see. In gases, the constituent particles are indistinguishable; in terms of quantum mechanics, each of them can be regarded as a wave that fills the entire container, with permitted energy states

$$E = \frac{1}{2m}\left(p_x^2 + p_y^2 + p_z^2\right)$$

where p_x, p_y, and p_z are the x, y, and z components of its momentum and L is the length of a side of the container (which is assumed to be a cube). The momenta take discrete values

$$p_x = hn_1/L, \qquad p_y = hn_2/L, \qquad p_z = hn_3/L$$

where h is Planck's constant and n_1, n_2, and n_3 can be any positive or negative integers, including zero. The magnitude of the momentum vector $|\mathbf{p}| = (p_x^2 + p_y^2 + p_z^2)^{1/2}$; $\mathbf{p} = \hbar\mathbf{k}$ where $\mathbf{k}$ is the wave vector.

Particles come in two kinds: (1) bosons, with symmetric wave functions, in which case any number of them can occupy the same quantum state (i.e., having the same values of n_1, n_2, and n_3); and (2) fermions, with antisymmetric wave functions such that there is a rigid restriction that (neglecting spin) precludes the occupation of a given state by more than one particle. Examples of bosons include photons, phonons, and entities, such as ^{4}He atoms, that are made up from an even number of fundamental particles. Examples of fermions include electrons, protons, neutrons, and entities, such as ^{3}He atoms, that are made up from an odd number of fundamental particles. Bosons are described by *Bose–Einstein statistics*; fermions by *Fermi–Dirac statistics*.

At extremely low temperatures, therefore, the particles of a boson assembly would almost all be in the same zero momentum state with $n_1 = n_2 = n_3 = 0$. A fermion assembly would also be in a state close to its lowest possible energy, but this would be very different from zero, since only one particle can occupy the zero mementum state and the others must therefore be in excited states of finite momentum and energy.

In practice, one deals with very large numbers of particles, the levels are very closely spaced, and it is reasonable to envisage a fermion system at very low temperatures as a sphere of filled states in momentum space as sketched in Fig. 2; that is, in a space whose axes represent the components of momenta in the x, y, and z directions. As the temperature is gradually raised, the surface of the sphere becomes "fuzzy" as a few particles just under the surface get promoted to unfilled states just above the surface. At higher temperatures the sphere progressively becomes less

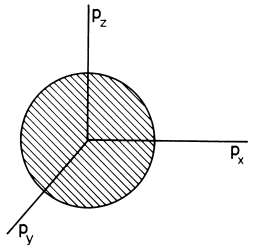


FIGURE 2 Sketch of the Fermi sphere of filled states in momentum space for an ideal gas of fermions at a temperature $T \ll T_F$. All states with momenta $p \leq (p_x^2 + p_y^2 + p_z^2)^{1/2}$ are filled. Those for larger momenta are all empty.

well defined and finally "evaporates." For high temperatures, the average energy of the particles is large enough that the likelihood of two of them wanting to be in the same state is negligible, since they are then well spread out in momentum space; it has consequently become irrelevant whether they are bosons or fermions, and their behavior is classical. The sphere of filled states at low temperatures is known as the *Fermi sphere*, and the particles at its surface, the *Fermi surface*, possess the *Fermi energy* E_F and the *Fermi momentum* p_F. The criterion as to whether the temperature T is to be regarded as "high" or "low" is whether $T > T_F$. or $T < T_F$, where $E_F = k_B T_F$, k_B is Boltzmann's constant, and

$$T_F = (3\pi^2 N/V)^{2/3}(h^2/4\pi^2 m k_B)$$

where N/V is the number of particles of mass m per unit volume. T_F is known as the *Fermi temperature*.

An interesting phenomenon predicted for boson systems where the number of particles is fixed is that of *Bose–Einstein condensation*. This implies that, as the temperature is gradually reduced. there is a *sudden* occupation of the zero momentum state by a macroscopic number of particles. The number of particles per increment range of energy is sketched in Fig. 3 for temperatures above and just below the Bose–Einstein condensation temperature T_b. The distribution retains, at least approximately. the classical Maxwellian form until T_b is reached. Below T_b there are two classes of particles: those in the condensate (represented by the "spike" at $E = 0$) and those in excited states. The criterion for a high or low temperature in a boson system is simply that of whether $T > T_b$ or $T < T_b$, where T_b is given by:

$$T_b = \left(\frac{N/V}{2.612}\right)^{2/3} \frac{h^2}{2\pi m k_B}$$

In the case of particles that are localized, such as paramagnetic ions within a crystal lattice, it is largely irrelevant whether they are bosons or fermions. They are distinguishable by virtue of their positions in the lattice and, for the

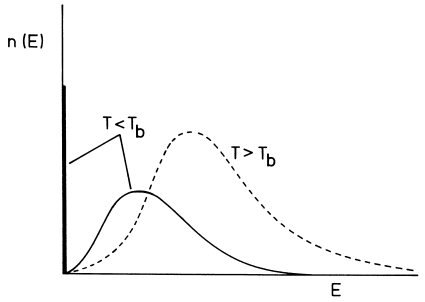


FIGURE 3 Sketch of the energy distribution function $n(E)$ of an ideal Bose–Einstein gas above, and very slightly below, the Bose–Einstein condensation temperature T_b.

same reason, there is no way in which two of them could ever be in the same state. Such assemblies are described by Boltzmann statistics. In a system containing N localized particles, the number n of them occupying a particular energy level E_i (each on its own separate site) is given by:

$$n = Ne^{-E_i/kT} \Big/ \sum_j e^{-E_j/kT}$$

where the summation is taken over all the permitted energy levels.

II. SOLIDS AT LOW TEMPERATURES

A. Insulators

The thermal energy in an insulating crystal takes the form of vibrations of the atoms about their equilibrium positions. The atoms do not vibrate independently, however, since the movements of one atom directly modulate the potential wells occupied by its neighbors. The result is coupled modes of vibration of the crystal as a whole, known as *phonons*. The phonon dispersion curve for an idealized one-dimensional lattice is sketched in Fig. 4. It is periodic, and it may be seen that there is a definite maximum phonon frequency, known as the Debye cut-off frequency ν_D. This corresponds physically to the fact that the shortest meaningful wavelength of a vibration in a discrete lattice is equal to twice the nearest neighbor separation; vibrations with apparently shorter wavelengths than this are indistinguishable from equivalent vibrations of longer wavelength. corresponding to the bending over of the dispersion curves at wave vectors of $\pm\pi/a$. It is also an expression of the fact there are only a limited number of normal modes of vibration of the crystal, equal to $3N$ where N is the total number of atoms in it. All the physical information about the dispersion curve is therefore contained in the range of wave vectors between $\pm\pi/a$, which is known as the *first Brillouin zone*. Dispersion curves for real three-dimensional solids are considerably more complicated than this, and there are longitudinal and transverse modes of vibration to be considered, but the essential physics is still as summarized in Fig. 4.

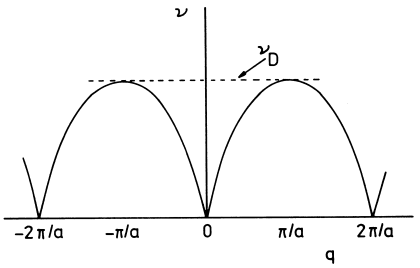


FIGURE 4 The phonon dispersion curve for an ideal one-dimensional lattice; the phonon frequency ν is plotted as a function of its wave vector **q**. The phonon energy $E = h\nu$.

Phonons in a crystal are closely analogous to the photons of electromagnetic radiation (although the latter do not, of course, have any known upper limit on their frequency). Just as in the case of photons, it is possible to envisage phonons as traveling, particle-like packets of energy. They move through the crystal at the velocity of sound (except for very high-energy phonons whose group velocities, given by the gradient of the dispersion curve, fall towards zero on the Brillouin zone boundary).

An excellent way of deducing the specific heat of a crystal is to model it as an empty box containing a gas of phonons. The phonons are bosons so Bose–Einstein statistics must be used. It should be noted, however, that the total number of phonons is not conserved (so that a Bose–Einstein condensation is not to be expected), quite unlike a gas of, for example, helium atoms. This approach yields the famous Debye specific heat curve shown in Fig. 5. It is plotted, not as a function of T directly, but as a function of T/θ_D where $\theta_D = h\nu_D/k_B$ is the Debye characteristic temperature, thereby yielding a universal curve that turns out to be in excellent agreement with experiment for a very wide range of materials. Whether or not the temperature should be considered to be low in the context of any given material depends on whether or not $T \ll \theta_D$. Representative values of θ_D for a few selected materials are given in Table I.

It can be seen from Fig. 5 that the specific heat approaches the classical (Dulong and Petit) value of $3R$ when $T \gg \theta_D$, as predicted by the theorem of the Equipartition of Energy. At lower temperatures, however, the specific heat decreases rapidly towards zero in accordance with the third law. For $T \ll \theta_D$, it is found, both experimentally and theoretically, that $C_v \propto T^3$.

The magnitude of the thermal conductivity coefficient κ of an insulator is determined both by the number of phonons it contains and also by the ease with which they

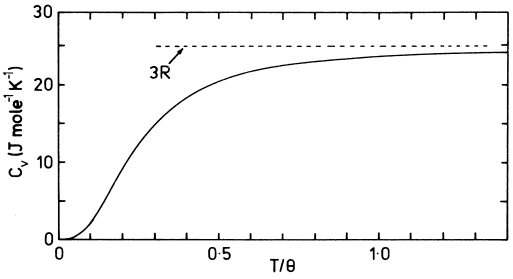


FIGURE 5 The molar specific heat C_v of an insulating solid, according to Debye theory, plotted as a function of T/θ_D where T is the absolute temperature and θ_D is the Debye characteristic temperature for any given material (see Table I).

can travel through the lattice. In fact, the process can accurately be modeled by analogy with heat conduction in a classical gas, for which

$$\kappa = C_V v \lambda / 3$$

where v is the velocity and λ the mean free path of the particles. In the present case, C_V is given by the Debye curve of Fig. 5, v is the (approximately constant) velocity of sound, and λ is determined by a variety of scattering processes. A typical $\kappa(T)$ curve for a crystal is shown by the upper curve of Fig. 6. At high temperatures, where C_V is constant, κ falls with increasing T because of *Umklapp processes*: a pair of phonons interacts to produce a third phonon of such large momentum that it lies outside the first Brillouin zone or, equivalently, lies within the first Brillouin zone but with a momentum of opposite sign from those of the initial phonons. The result is that the group velocity and direction of energy flow are reversed. Umklapp processes are very effective in limiting the thermal conductivity at high temperatures. As the temperature is reduced, however, there are fewer and fewer phonons of sufficient energy for the process to occur, so λ increases rapidly and so does κ. Eventually, for a good-quality single crystal, λ becomes equal to the dimensions of the sample and (usually) can increase no further; the mean free path is then limited by *boundary scattering* in which the phonons are scattered by the walls and not to any important extent by each other. In this regime, λ is constant and approximately equal to the diameter of the crystal, v is constant, and C_V is decreasing with falling T since $T < \theta_D$ (see Fig. 5), so that κ passes through a maximum and also decreases, varying at T^3, like C_V, in the low temperature limit.

There are many other scattering processes that also influence κ. Phonons may, for example, be scattered by impurities, defects, and grain boundaries. Furthermore, boundary scattering at low temperatures can be either specular (equal angles of incidence and reflection) or, more commonly, diffuse (random angle of reflection). The former process does not limit λ, which can consequently become considerably larger than the diameter of the crystal; it occurs when the phonon wavelength is much longer than the scale of the surface roughness, for example in the

TABLE I Debye Characteristic Temperatures θ_D for Some Common Elements[a]

Element	θ_D(K)
Ag	225
Al	426
Au	164
C (diamond)	2065
Co	443
Cr	585
Fe	464
H (para)	116
He[b]	28–36
Hg	75
N	68
Ni	440
O	91
Pb	108
Si	636
Sn[c]	212
Zn	310

[a] From Rosenberg. H. M. (1963). "Low Temperature Solid State Physics," Clarendon Press, Oxford.
[b] Precise value dependent on pressure.
[c] Value given for grey tin.

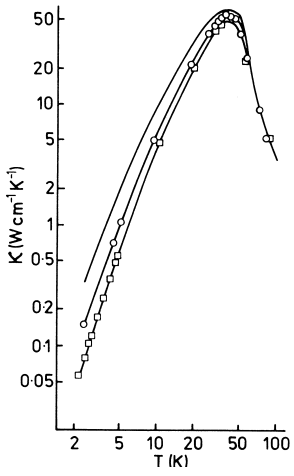


FIGURE 6 The thermal conductivity κ of a synthetic sapphire crystal of diameter 3 mm (upper curve), plotted as a function of temperature T. The circles and the associated curve represent measurements of κ for the same sample after it had been ground down to a diameter of 1.55 mm; and the squares, a diameter of 1.02 mm. [From Berman, R., Foster, E. L., and Ziman, J. M. (1955). *Proc. R. Soc. A* **231,** 130.]

case of a flame-polished sapphire. Thermal conduction in the boundary scattering regime is strange, at first sight, in that the magnitude of κ is geometry and size dependent, becoming smaller if the physical dimensions of the sample are reduced.

The addition of impurities to a crystal drastically reduces κ by providing an additional phonon scattering mechanism, thereby reducing λ. Similarly, κ for disordered or glassy insulators is usually very small because of the greatly enhanced phonon scattering.

B. Metals

In relation to its low temperature properties, a metal can be viewed as being like the insulating crystal discussed above but where, in addition, there is a gas of highly mobile quasi-free conduction electrons. The lattice has remarkably little effect on their motion, apart from opening up energy gaps, or forbidden energy ranges, at Brillouin zone boundaries. The electrons can be treated as an ideal Fermi gas.

Measurements of the specific heats of metals yield results almost indistinguishable from the Debye specific heat curve of Fig. 5, and it can be concluded, therefore, that the electrons make a negligible contribution to the specific heat. This is at first sight astonishing since, according to the classical Equipartition of Energy theorem, there should on average be a thermal energy of $\frac{3}{2}k_{\mathrm{B}}T$ per electron ($\frac{1}{2}k_{\mathrm{B}}T$ per degree of freedom, of which there are three for a gas), leading to an electronic molar specific heat contribution of $\frac{3}{2}$R. The result is easily explicable, however, in terms of Fermi–Dirac statistics. The Fermi temperature of the electron gas in a typical metal turns out to be $T_{\mathrm{F}} \simeq 5 \times 10^4$ K. Thus, even at room temperature, $T \ll T_{\mathrm{F}}$, and the electrons must be regarded as being effectively at a very low temperature. There is consequently a well-defined Fermi sphere in momentum space (see Fig. 2). Most of the electrons, being deep inside the sphere, cannot change state because there are no empty adjacent states into which to move and so are unresponsive to an alteration of temperature and do not contribute to the specific heat. The only electrons that contribute are those within an energy range of $\sim \pm\, k_{\mathrm{B}}T$ of the Fermi surface, which can be shown to lead to a linear dependence of C_V on T. Such behavior is characteristic of highly degenerate ($T \ll T_{\mathrm{F}}$) Fermi systems and is also seen in liquid ^{3}He and in liquid ^{3}He–^{4}He solutions (see below).

With a lattice (phonon) specific heat varying as T^3 at low temperatures, and an electron specific heat varying as T, it is evident that the latter contribution will eventually dominate if the temperature is reduced sufficiently. For a typical metal, it is necessary to go down at least to temperatures near 1 K in order to measure the electronic contribution to the specific heat.

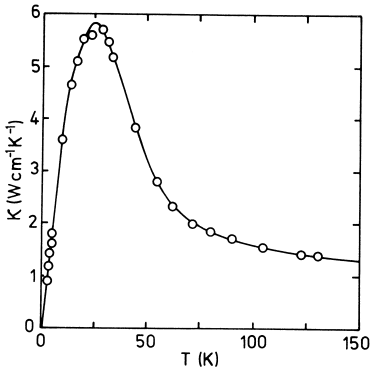


FIGURE 7 The thermal conductivity κ of chromium plotted as a function of temperature T. [From Harper. A. F. A., Kemp, W. R. G., Klements, P. G., Tainsh, R. J., and White, G. K. (1957). *Philos. Mag*. **2,** 577.]

Thermal conduction in metals can in principle occur through both the phonon and the electron gases. However, the phonons tend to scatter the electrons, and *vice versa*; in practice, it turns out that most of the conduction is due to the electrons. A typical $\kappa(T)$ characteristic for a pure metal is shown in Fig. 7. At high temperatures electron-phonon scattering limits the electronic mean free path, so κ rises with decreasing T because of the corresponding decrease in the number (and average energy) of phonons. Eventually, the phonons become so few in number that this scattering process becomes unimportant. The electronic mean free path then becomes temperature independent, being limited by defect scattering or, for a very pure and perfect crystal, by boundary scattering. The heat capacity falls linearly with decreasing T; the relevant velocity for the electrons is the Fermi velocity v_{F}, which is constant; the mean free path is constant; and κ, too, falls linearly with T at very low temperatures.

The electron gas can also, of course, support the passage of an electrical current. The electrical resistivity ρ is governed by electron–phonon scattering at high temperatures. It consequently falls as T is decreased; a typical result is as sketched in Fig. 8. At a sufficiently low temperature, electron–phonon scattering becomes negligible and the resistivity approaches a temperature-independent value, ρ_0, known as the *residual resistivity*. The magnitude of ρ_0 provides an excellent measure of the purity and crystalline quality of the sample and is usually quoted in terms of the *resistivity ratio*, the ratio by which the resistivity changes on cooling from room temperature. In extreme cases of purity and crystal perfection, resistivity ratios as large as 10^5 have been recorded.

Both the electrical and thermal conductivities of metals are greatly reduced by impurities or imperfections in the lattice. Alloys, for example, are an extreme case and usually have conductivities that are many orders of magnitude

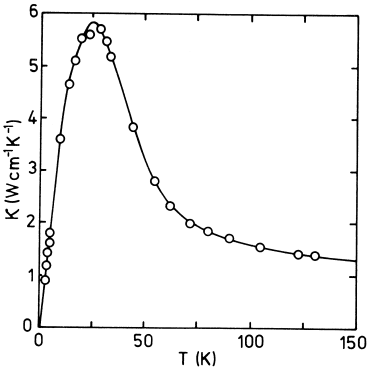


FIGURE 8 Sketch to indicate the variation of electrical resistivity ρ with temperature T in a metal. In the limit of very low temperatures, ρ usually approaches a constant value, ρ_0, known as the *residual resistivity*.

lower than those of pure metals. For this reason, they are especially valuable in the construction of cryostats since less heat will then flow from the hotter to the colder parts of the system. Stainless steel and, to a lesser extent, brass and German silver, are frequently used for such purposes.

C. Superconductivity

There are a large number of materials whose electrical resistance vanishes abruptly at a critical temperature T_c somewhere in the range below 120 K, corresponding to the onset of super conductivity. The phenomenon is a remarkable one, for which no known analog exists at room temperature. The transition itself is extremely sharp for a pure metal (with a width of 10^{-5} K in gallium, for example), and the electrical resistance at very low temperatures does actually appear to be zero. A circulating current in a closed loop of the superconductor, which is easily detectable by the external magnetic field that it creates, will flow without measurable diminution for as long as the experiment is continued.

As well as being perfect conductors, superconductors are also perfect diamagnets. An applied magnetic field is always excluded from the bulk of the material, regardless of whether the field was applied before or after cooling through the superconducting transition at T_c. This phenomenon, known as the *Meissner effect*, involves the appearance of circulating currents at the surface of the superconductor which create a magnetic field that exactly cancels the applied one, inside the material.

The basic mechanism of superconductivity was accounted for in considerable detail by the theory of Bardeen, Cooper, and Schrieffer (BCS theory), which discusses the phenomenon in terms of the formation of *Cooper pairs* of electrons through an attractive interaction mediated by the lattice. Unlike individual electrons, the Cooper pairs are bosons and form a condensate analogous to those for liquid ^{4}He or liquid ^{3}He below their superfluid transition temperatures. The condensate possesses a macroscopic wave function that extends throughout the sample and implies the coherent collective motion of a very large number of particles. To change the state of any single particle would involve making simulataneous changes in the states of all the others, which is a highly improbable event and helps to explain the resistanceless flow of a current.

III. LIQUID HELIUM

A. Liquid ^{4}He and Liquid ^{3}He

The low-temperature phase diagrams of ^{4}He and ^{3}He are shown in Fig. 9(a) and (b), drawn on the same scale for convenient comparison. It can be seen immediately that, although the two phase diagrams clearly have much in common, there are also some interesting differences. Both isotopes remain liquid to the lowest attainable temperatures, but each can be solidified by the application of sufficient external pressure, greater pressure being needed in the case of ^{3}He (35 bar in the low-temperature limit) than in the case of ^{4}He (25 bar). A pronounced minimum is evident at ~0.3 K in the melting curve of ^{3}He. Both liquids undergo superfluid transitions, although, as remarked above, that for ^{3}He takes place at a temperature (T_c) a thousand times lower than that (T_λ) for ^{4}He.

The non-solidification of the liquid heliums at very low temperatures is, in part, a consequence of the extreme weakness of their interatomic forces. The helium atom

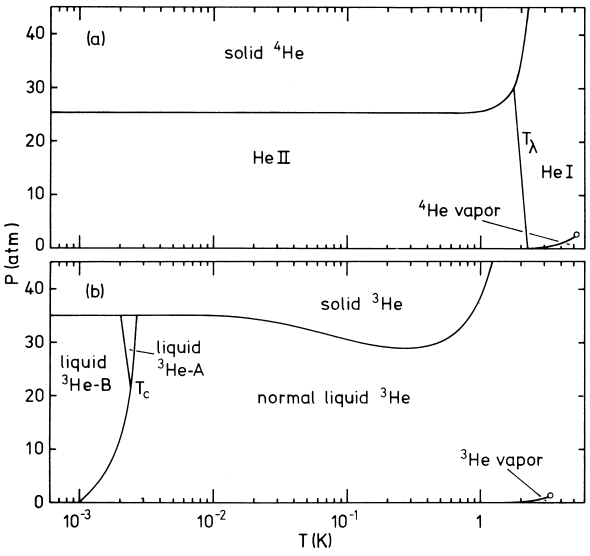


FIGURE 9 The low-temperature phase diagrams of (a) ^{4}He and (b) ^{3}He in zero magnetic field.

possesses an inert gas structure and is chemically inert. It can be envisaged as a small, rigid sphere, much like the hard sphere conventionally assumed by classical kinetic theory. The only attractive force that exists between helium atoms is that due to the Van der Waals interaction, which is exceptionally weak owing to the tightness with which the two electrons are bound to the nucleus (helium having the highest ionization potential of any atom). The weakness of the interatomic forces would not, in itself, prevent the solidification of helium under its saturated vapor pressure, though it would imply, of course, that a very low temperature would be needed for the process to occur.

The other vital consideration affecting the possibility of solidification and influencing also almost all of the other properties of liquid helium is its very high zero-point energy. In quantum mechanics, the lowest permitted kinetic energy for a particle is not necessarily zero. If a particle of mass m is confined to a volume V, then its minimum kinetic energy, or zero-point energy, is

$$E_z = h^2/8m(4\pi/3V)^{2/3}$$

where h is Planck's constant. The zero point energy of helium is particularly large because of its relatively small atomic mass.

In order to minimize its total energy, the liquid tends to expand, thereby increasing V and decreasing E_z. Because of the weakness of the interatomic forces, the effect of zero point energy on the density of the liquid is very large. As a result, liquid ^{4}He has a density only one-third of the value that would be obtained if the atoms were touching each other. Liquid ^{3}He is affected even more strongly on account of its smaller atomic mass, its density being reduced by a factor of about one-quarter. Both isotopes of helium can thus be envisaged as forming liquids in which the atoms are exceptionally widely spaced and therefore able to slip past each other particularly easily; liquid helium, therefore, has rather gas-like properties.

The key to an understanding of the liquid heliums lies, in fact, in the realization that they are often best considered as though they were gases. As a first approximation, therefore, they can be treated as ideal gases. Liquid ^{4}He, being composed of atoms that are bosons, should thus be subject to Bose–Einstein statistics; liquid ^{3}He, composed of fermions, should be subject to Fermi–Dirac statistics. Any deviations of the properties of the actual liquids from those predicted for ideal gases of the same density can then be attributed to the finite sizes of the atoms and to the attractive interactions between them.

At temperatures above 2.17 K, liquid ^{3}He and liquid ^{4}He are very similar to each other and both behave like rather non-ideal dense classical gases. At 2.17 K under its natural vapor pressure, however, ^{4}He undergoes the so-called lambda transition and acquires a whole new set of properties, many of which are quite unexpected in character.

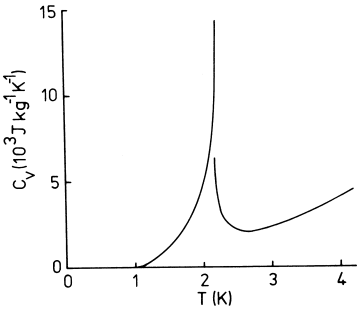


FIGURE 10 The specific heat C_V of liquid ^{4}He as a function of temperature T [From Atkins, K. R. (1959). "Liquid Helium." Cambridge Univ. Press, London.

B. Superfluid Properties of Liquid ^{4}He

The properties of liquid ^{4}He above and below the lambda transition temperature T_λ are so completely different that it is almost as though one were dealing with two separate liquids; to emphasize the distinction they are referred to as He I and He II respectively. The name of the transition derives from the characteristic shape of the logarithmic singularity that occurs in the specific heat (Fig. 10) at that temperature, which is strongly reminiscent of a Greek lambda. As can be seen from Fig. 9, T_λ is weakly pressure dependent, falling towards lower temperatures as the pressure is increased.

One of the most striking properties of He II is its ability to flow effortlessly through very narrow channels. An interesting demonstration can be seen in superfluid film flow, illustrated in Fig. 11, where a vessel containing He II tends to empty; and an empty one that is partially immersed in He II tends to fill up until the inside and outside levels

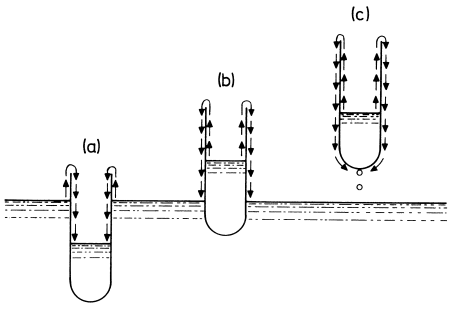


FIGURE 11 Gravitational flow of the creeping superfluid film. (a) and (b) In the case of a partially immersed vessel, film flow equalizes the inside and outside levels. (c) A suspended vessel of He II will eventually empty completely via the film. [From Daunt, J. G., and Mendelssohn, K. (1939). *Proc. R. Soc.* (*London*) **A170,** 423.]

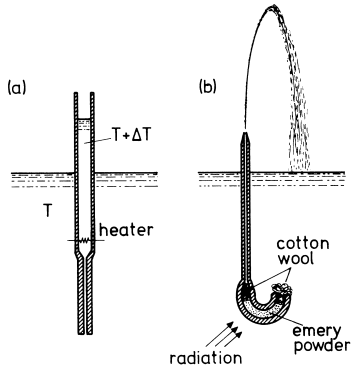


FIGURE 12 (a) When the temperature of the He II inside the vessel is raised slightly above that of the bath, liquid flows in through the capillary tube, which acts as a superleak. (b) The same phenomenon (using radiative heating in the case illustrated) can be used to create a dramatic fountain of liquid helium. [From Allen, J. F., and Jones, H. (1938). *Nature* **141,** 243.]

equalize. What happens is that the adsorbed film of helium on the walls of the vessel acts as a syphon through which the liquid can pass under the influence of the gravitational potential. The same phenomenon would doubtless take place for other liquids, too, such as water, were it not that their viscosities effectively immobilize their films on the walls, so that flow at an appreciable rate cannot take place. In the case of He II, rapid nonviscous flow occurs within the film which, despite its narrowness ($\sim$10 nm), can still carry an appreciable volume flow rate of liquid.

It would be a gross oversimplification, however, to say that He II is inviscid. The situation is much more complicated than this, as can be seen from the fact there are often thermal effects associated with superflow. For example, when He II is allowed to drain partially from a vessel whose exit is blocked with a superleak of tightly packed power (impenetrable to conventional fluids) it is found that the remaining liquid is at a higher temperature than it was at the start. The inverse experiment can also be carried out. When an open-topped cylinder is connected to a reservoir of He II via a superleak (e.g., a very fine capillary tube) as shown in Fig. 12a, and the He II inside the cylinder is warmed to a temperature slightly above that of the reservoir, its level rises. If the top of the vessel is drawn out into a fine jet and the temperature of the He II inside is raised sufficiently, the liquid squirts to form a fountain as shown in Fig. 12(b), hence the common apellation *fountain effect* for this phenomenon.

Attempts to measure the viscosity of He II directly can lead to ambiguous results. The liquid can flow through a narrow tube or orifice, with zero pressure head needed to induce the flow, implying that the viscosity is itself zero. On the other hand, finite viscosity values can be measured through investigations of the damping of oscillating disks or spheres immersed in the liquid.

Two seemingly quite different models have been developed to describe the strange modes of behavior of He II: the *two-fluid model* and the *excitation model*, each of which has its own range of utility and appliability. It can be demonstrated theoretically that the two pictures are formally equivalent, in that the latter model implies the former.

We discuss first the phenomenological two-fluid model, which is generally the more useful in describing the behavior of the liquid between 0.5 K and T_λ. The model postulates that He II consists of an intimate mixture of two separate but freely interpenetrating fluids, each of which fills the entire container. There is a superfluid component of density ρ_s that has zero viscosity and carries no entropy, and there is a normal fluid component of density ρ_n that has a finite viscosity, carries the whole entropy of the liquid, and behaves much like any ordinary fluid. The ratios ρ_s/ρ and ρ_n/ρ are not predicted by the model, but can readily be determined by experiment, when it is found that $\rho_s/\rho \to 1$ as $T \to 0$ and $\rho_s/\rho \to 0$ as $T \to T_\lambda$ (with opposite behavior occurring in ρ_n/ρ so as to keep the total density ρ equal to $\rho_n + \rho_s$), as shown in Fig. 13. Any transfer of heat to the liquid transforms some of the superfluid into normal fluid in such a way that the total density $(\rho_n + \rho_s)$ remains constant. A form of two-fluid hydrodynamics, taking explicit account of these thermal effects, has been developed and gives an excellent description of most of the properties of the liquid. It is found that, unlike ordinary fluids, the superfluid may be accelerated, not only by a gradient in pressure but also by a gradient in temperature.

The two-fluid model provides a conceptual framework within which most of the phenomena associated with He II can be understood. For example, a viscosity of zero is found for flow-through fine channels, because only the superfluid component then moves. When a sphere or disk oscillates in bulk He II, however, its motion is damped by the normal fluid component which surrounds it, and the

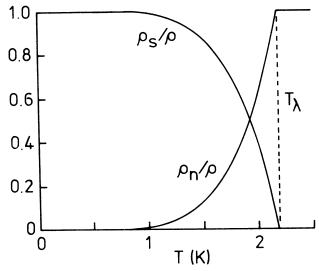


FIGURE 13 The normal and superfluid densities, ρ_n and ρ_s, of He II as functions of temperature T, divided in each case by the total density ρ of the liquid.

viscosity appears to be finite. The rise in the temperature of the residual liquid observed when He II is allowed to drain away through a superleak occurs because only the superfluid component leaves so the entropy of the liquid left behind must remain constant; its entropy density, and hence its temperature, therefore increases.

The liquid is effectively a superconductor of heat, so it is extremely difficult to create a temperature gradient. If a heater is switched on in the liquid, it converts superfluid component to normal fluid, which flows away while more superfluid moves through it towards the heater—an extremely efficient heat-removal process. This *superfluid heat flush* effect carries all suspended impurities away from the heater, because they are part of the normal fluid, and it can therefore be used as the basis of a technique for isotopic purification.

New phenomena have been predicted on the basis of the two-fluid equations of motion—for example, the existence of a novel wave mode known as *second sound* in which the normal and superfluid components oscillate in antiphase such that the density remains constant. The characteristic wave velocity (typically 20 m sec^{-1} but depending on temperature) is much slower than that of ordinary or first sound (typically 240 m sec^{-1}), a pressure-density wave in which the two components oscillate in phase. Second sound is a temperature-entropy wave at constant pressure and density; it may be created by means of an oscillating voltage applied to a heater and may be detected by means of a thermometer.

It is tempting to identify the superfluid and normal fluid components of He II with the atoms of the condensate and excited states, respectively, of an ideal Bose–Einstein gas, but this, too, turns out to be a gross oversimplification. Nonetheless, it is clear that Bose–Einstein statistics do play an essential role in determining the properties of the liquid. If one calculates the Bose–Einstein condensation temperature of an ideal gas with the same density and atomic mass as liquid ^{4}He, one finds $T_c \simeq 3.1$ K, and the difference between this value and T_λ may plausibly be ascribed to the influence of the interatomic forces. The presence of a condensate of ^{4}He atoms in the zero momentum state has recently been confirmed by inelastic neutron scattering measurements. As shown in Fig. 14, the condensate fraction rises from zero at T_λ towards an asymptotic limit of about 14% at low temperatures. This departure from the approximately 100% condensate characteristic of an ideal gas for $T \ll T_c$ is an expected consequence of the non-negligible interatomic forces. Of couse, as already noted in Fig. 13, the superfluid component does form 100% of the liquid at $T = 0$ and it cannot, therefore, be directly identified with the condensate fraction.

The excitation model of He II perceives the liquid as being somewhat akin to a crystal at low temperatures. That

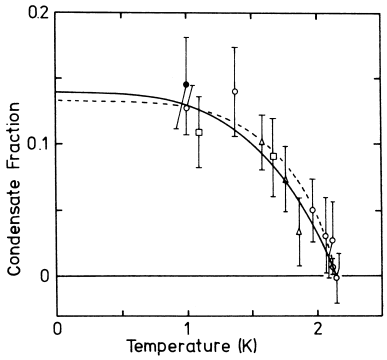


FIGURE 14 The Bose–Einstein condensate fraction of liquid ^{4}He as a function of temperature T. The points and the dashed curve represent experimental data from various sources. The full curve is a theoretical temperature dependence fitted to the data, enbling a $T = 0$ K value of $14 \pm 2\%$ to be deduced. [From Sears, V. F., Svensson, E. C., Martel, P., and Woods, A. D. B. (1982). *Phys. Rev. Lett.* **49,** 279; Campbell, L. J. (1983). *Phys. Rev. B* **27,** 1913.]

is, the liquid is envisaged as an inert background "either" in which there can move a gas of particle-like excitations which carry all the entropy of the system. The dispersion curve of the excitations as determined by inelastic neutron scattering is shown in Fig. 15. Usually, only those excitations within the thickened portions of the curve need be considered, corresponding to *phonons* at low momentum and to *rotons* at higher momentum. The phonons are quantized, longituadinal sound waves, much like those found in crystals; the physical nature of the rotons remains something of an enigma.

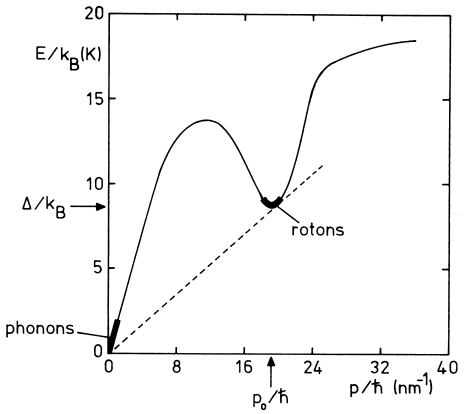


FIGURE 15 The dispersion curve for elementary excitations in He II under its saturated vapor pressure. The energy E of an excitation (divided by Boltzmann's constant to yield temperature units) is plotted against its momentum p (divided by h to yield the wave vector in units of reciprocal length). Only those states within the thickened portions of the curve, *phonons* and *rotons*, are populated to any significant extent in thermal equilibrium. The graident of the dashed line drawn from the origin to make a tangent near the roton minimum is equal to the Landau critical velocity v_L. [From Cowley, R. A., and Woods, A. D. B. (1971). *Can. J. Phys*. **49,** 177.]

The entropy and specific heat of the liquid are readily calculated by standard statistical mechanical methods in terms of the shape of the dispersion curve; that is, in terms of the velocity of sound c, which specifies the gradient dE/dP at small values of p, and of the roton parameters Δ, p_0, and μ, which specify, respectively, the energy, momentum, and effective mass of a roton at the minimum. Good agreement is obtained between calculated and measured values, provided that the temperature is not too high. Close to T_λ, this simple picture breaks down, for two reasons. First, the roton parameters themselves become temperature dependent, and, second, the short lifetimes of the excitations cause a broadening of their energies. Again, provided that the temperature is not too high, it is possible to calculate the momentum carried by the gas of excitations. This quantity may be identified with the momentum carried by the normal fluid component of the two-fluid model. Indeed, it emerges that, in a very real sense, the excitations *are* the normal fluid component. It is then straightforward to calculate ρ_s and ρ_n as functions of T, in terms of c, Δ, p_0, and μ.

The excitation model also provides a natural explanation for the superfluidity of He II. If it is assumed that a moving object in the liquid can dissipate energy only through the creation of excitations of energy E and momentum p and that energy and momentum must be locally conserved, then it can easily be demonstrated that the minimum velocity at which such processes can occur is given by E/p. For most liquids, where the thermal energy is associated with individual atoms or molecules, $E = p^2/2m$ and E/p can take any value down to zero. The same argument would apply to the Bose–condensate for an ideal gas. In the case of He II, however, E/p can never fall below the value given by the gradient of a tangent drawn from the origin to the roton minimum of the dispersion curve, as shown in Fig. 15. This is the minimum velocity at which dissipation can occur in the superfluid component and is known as the Landau critical velocity, v_L. Values of v_L have been determined from measurements of the drag on a small moving object (a negative ion) in He II as a function of its speed v. As shown in Fig. 16, the results show that there is no drag on the object until a speed of v_L has been attained, in excellent agreement with theoretical predictions based on the dispersion curve.

Fundamental (quantum mechanical, manybody) theories of He II usually seek to calculate the dispersion curve of Fig. 15 starting from a knowledge of the radii and massess of ^{4}He atoms and of the interactions between them. Once the dispersion curve has been correctly derived, superfluidity and the two-fluid model follow automatically.

In actual fact, critical velocities measured for He II, for example in flow experiments, are usually a great deal

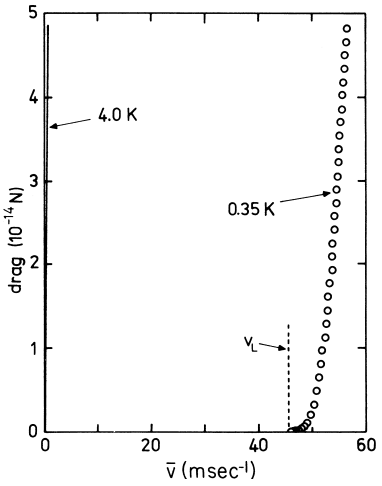


FIGURE 16 The drag force on a negative ion moving through He II at 0.35 K, as a function of its speed $\bar{v}$. The drag remains immeasurably small until $\bar{v}$ has reached the Landau critical velocity for roton creation, v_L, but then rises rapidly with further increase of $\bar{v}$. The behavior of a negative ion in He I at 4.0 K is also plotted, in order to emphasize the profound qualitative difference between the two cases. [From Allum, D. R., McClintock, P. V. E., Phillips, A., and Bowley, R. M. (1977). *Phil. Trans. R. Soc. (London)* **A284,** 179.]

smaller than v_L. This is not because there is anything wrong with the physical arguments leading to v_L; rather, it is an indication that other kinds of excitation can exist in the liquid, in addition to phonons, rotons, and related regions of the dispersion curve. In particular, *quantized vortices* can exist in the superfluid component. A quantized vortex in He II is a linear singularity, with a core region of atomic or subatomic dimensions, around which the superfluid flows tangentially with a velocity $\mathbf{v}_s$, that varies inversely with radial distance from the core, so that the circulation:

$$\kappa_c = \int \mathbf{v}_s \cdot \mathbf{dl} = nh/m_4$$

where d is a superfluid order parameter, l the intrinsic angular momentum, h Planck's constant, m_4 the mass of a ^{4}He atom, and n an integer. In practice, for free vortices n is always equal to unity. Samples of He II, regardless of their history, always seem to contain at least a few quantized vortices. When a characteristic critical flow velocity is exceeded (typically of a few mm sec^{-1} or cm sec^{-1}, depending on the channel geometry and dimensions), these grow into a dense disordered tangle, somewhat akin to a mass of spaghetti but with no free ends, thereby giving rise to a dissipation of the kineitc energy of the flowing liquid.

Quantized vortices also play a central role in the "rotation" of the He II within a rotating vessel. Strictly, the

superfluid component of He II cannot rotate at all because of a requirement that:

$$\text{curl } \mathbf{v}_s = 0$$

which, by use of Stokes' theorem, quickly leads to the conclusion that the circulation $\kappa_c = 0$ in any simply connected geometry (i.e., where any closed loop can in principle be shrunk to a point without intercepting boundaries). This condition implies that the superfluid component of He II in a vessel on the earth's surface does not rotate with the vessel but remains at rest relative to the fixed stars. There is convincing experimental evidence that this is indeed the case for small enough angular velocites. For sufficiently large angular velocities of the container, though, both components of the He II *seem* to come into rotation with the vessel. What is actually happening, however, is that a number of quantized vortices are formed parallel to the axis of rotation of the vessel; these cause the superfluid component to simulate solid-body-like rotation, because each vortex core moves in the combined flow fields of all the other vortices. Once vortices have been formed, the geometry is no longer simply connected, and the form of the flow field of a vortex is such that, except in the vortex cores, the liquid still remains curl-free, as required.

The inability of the superfluid component to rotate in a conventional manner is one of the most striking demonstrations of a macroscopic quantum phenomenon. The quantization condition for the vortices (above) emerges naturally from the postualte that the superfluid must be described by a *macroscopic wave function*, exactly analogous to the microscopic wave functions used to describe electrons in atoms, but on the scale of the containing vessel. The existence of macroscopic quantized vortices may seem surprising, but there is overwhelming evidence for their reality, culminating in the development of an imaging system that achieved the vortex photographs shown in Fig. 17.

Although quantized vortices are central to an understanding of He II, and most of their properties are now reasonably well understood, the mechanism by which they are created in the first place remains something of a mystery. As mentioned above, all samples of He II invariably seem to contain a few vortices. and it is now believed that these "background vortices" are formed as the liquid is being cooled through the lambda transition. It is also clear that vortices can be created in the superfluid at lower temperatures (for example, by a moving object), provided that a characteristic critical velocity is exceeded, but the mechanism is difficult to study because of the interfering effect of the background vortices.

The only reliable experiments on the *creation* of quantized vortices in the cold superfluid are those based on the use of negative ions. They have revealed that the mechanism involves a form of *macroscoopic quantum tunneling* (MQT); the whole system undergoes a discontinuous quantum transition, though an energy barrier, to a state with a vortex loop of finite length attached to the ion (there being no intermediate states for which momentum and energy can be conserved). Figure 18 shows measurements of the rate ν at which negative ions nucleate vortices in isotopically pure He II for three electric fields, as a function of reciprocal temperature T^{-1}. The flat regions of the curves on the right-hand side of the figure correspond to MQT through the energy barrier; the rapidly rising curves at higher temperatures (left-hand side) correspond to thermally activated jumps over the top of the barrier attributable to phonons.

Except for low velocities (usually much smaller than v_L) and small heat currents, the flow and thermal properties of He II are dominated by quantized vortices. If

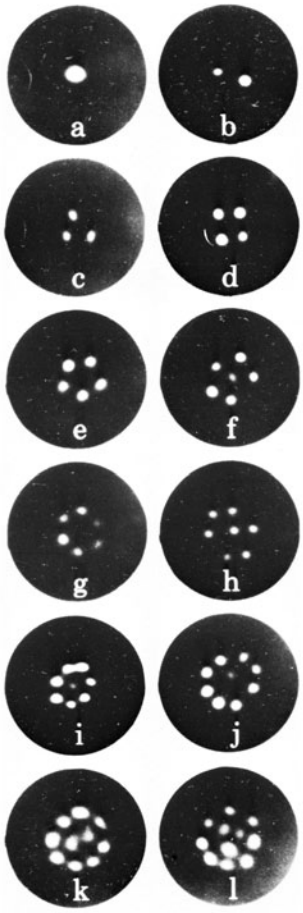


FIGURE 17 Quantized vortex lines in a rotating bucket of He II for various angular velocities, photographed from above. [From Yarmchuk, E. J., Gordon, M. J. V., and Packard. R. E. (1979). *Phys. Rev. Lett.* **43,** 214. Photograph by courtesy of R. E. Packard.]

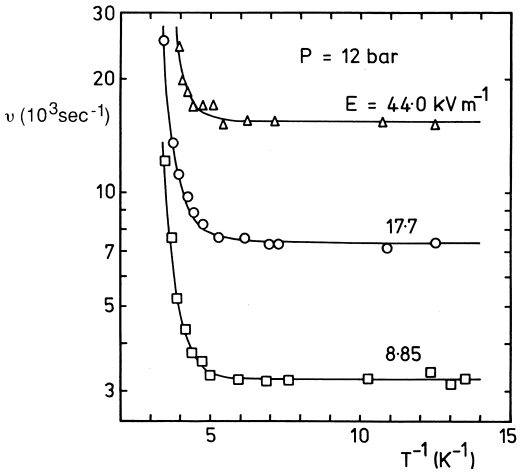


FIGURE 18 Measurements of the rate ν at which quantized vortices are created by negative ions moving through isotopically pure He II, plotted as a function of reciprocal temperature T^{-1} for three electic fields E. In stronger fields, the ions move faster and ν is correspondingly larger. The temperature-independent mechanism operative at low temperatures-independent mechanism operative at low temperatures (right-hand side of figure) is attributable to macroscopic quantum tunneling through an energy barrier; the rapid rise of ν with temperature at higher temperatures (left-hand side) is attributable to thermal activation over the barrier. [From Hendry, P. C., Lawson, N. S., McClintock, P. V. E., and Williams, C. D. H. (1988). *Phys. Rev. Lett.* **60,** 604.]

a critical velocity or heat current is exceeded, then the superfluidity breaks down and a self-sustained tangle of vortex lines builds up within the liquid. Being a metastable state, it continuously decays and can only be maintained by drawing energy from the flow field or heat current. Recent studies have shown that, despite its quantization, such turbulence can often behave in a surprisingly classical manner.

Quantized vortices in He II are the analog of the quantized flux lines found in superconductors, where magnetic field plays the same role as angular velocity in He II. From this viewpoint, He II can be regarded as analogous to an extreme Type II semiconductor. The very small core dimensions of the helium vortices implies, however, that the equivalent of the "second critical field" can never be attained in practice. Corresponding to that angular velocity of the system at which the vortex cores touched each other, it would require angular velocities far beyond the strengths of available container materials.

C. Normal Liquid ^{3}He

Liquid ^{3}He above 1 K is much like a dense classical gas and has properties very similar to those of He I. On further cooling, however, there is a *gradual* change in its behavior to a quite different set of properties, even for temperatures well above that of the superfluid transition, $T_c \simeq 1$ mK. Above 1 K, for example, the viscosity of the liquid is almost temperature independent and its thermal conductivity κ decreases with decreasing temperature, but at 50 mK, on the other hand, η varies as T^{-2} and κ as T^{-1}.

Qualitatively, this is very much the kind of behavior to be expected of an ideal Fermi gas. The Fermi temperature calculated for an ideal gas of the same density and atomic mass as liquid ^{3}He is $T_F \simeq 5$ K, but one has to bear in mind that a consequence of the interactions between the atoms is to increase the effective mass of each atom, thereby reducing T_F. The gradual change in properties of the liquid ^{3}He on cooling can therefore be attributed to the change between classical and Fermi–Dirac behavior as the liquid enters the degenerate ($T \ll T_F$) regime where a well-defined Fermi-sphere exists in momentum space. The specific heat of liquid ^{3}He becomes linear in T at very low temperatures, just as expected for an ideal Fermi gas, and is thus closely analogous to the specific heat of the electron gas in a metal. The paramagnetic spin susceptibility of the liquid, which varies as T^{-1} at high temperatures, becomes temperature independent below $\sim$100 mK. Again, this parallels the properties of the electron gas in a normal metal and is the behavior predicted theoretically for the ideal Fermi gas.

Although the properties of liquid ^{3}He are qualitatively almost identical with those of the ideal Fermi gas, it is not surprising that there are large quantitative differences; the system is, after all, a liquid and not a gas, so the interatomic forces cannot be left completely out of account. As already remarked, one effect of the interatomic forces is to raise the effective mass from m_3 for a bare ^{3}He atom to a larger value, m_3^*. It is not, however, possible on this basis alone to reconcile the departures from ideal behavior for all the properties of ^{3}He. That is to say, the values of m_3^* derived on this basis from measurements of different properties would themselves be widely different.

A detailed quantitative understanding of liquid ^{3}He at low temperatures requires the application of Landau's *Fermi liquid theory*, which takes explicit account of the interactions, and parameterises them in the form of a small number of dimensionless constants known as *Landau parameters*. For most purposes, only three of these parameters are needed (usually written as F_0, F_1 and G_1) and almost all of the properties of the interacting Fermi liquid can be calculated in terms of them. Numerical values of the Landau parameters are not predicted by the theory but are to be found by experiment. The crucial test of the theory—a requirement that consistent values of the Landau parameters should be obtained from widely differing kinds of experiment—is convincingly fulfilled.

An interesting situation arises in a Fermi system at very low temperatures when the average time τ between

collisions becomes very long. For any given sound frequency ω, there will be a temperature at which ω and τ^{-1} become equal. At lower temperatures than this, where $\omega\tau \gg 1$, the propagation of ordinary sound in an ideal Fermi gas clearly becomes impossible because the collisions simply do not occur fast enough. For the interacting Fermi liquid, however, Landau predicted that a novel, "collisionless" form of sound propagation known as *zero sound* should still be possible corresponding to a characteristic oscillating distortion in the shape of the Fermi sphere; zero sound has indeed been observed in liquid ^{3}He.

The curious minimum in the melting curve of ^{3}He near 0.3 K (Fig. 9) may be understood in terms of relative entropies of the two phases and of the Clausius–Clapeyron equation:

$$(dp/dT)_m = \Delta S_m/\Delta V_m$$

where m the subscript refers to the melting curve and ΔS and ΔV are the changes in entropy and volume that occur on melting. The negative sign of $(dP/dT)_m$ below 0.3 K (Fig. 9) shows that the solid entropy must then be *larger* than that of the liquid. In fact, this is only to be expected. Atoms in solid are localized on lattice sites and therefore distinguishable, so they will be described by Boltzmann statistics. The entropy of the solid below 1 K arises almost entirely from the nuclear spins, which can exist in either of two permitted states, and will therefore take the (temperature-independent) value of R ln 2 per mole in accordance with the Boltzmann–Planck equation. Given that the specific heat of the liquid, and therefore its entropy, varies approximately linearly with T, there is bound to be a characteristic temperature where the entropy of the liquid falls below that of the solid. Experimental values of the solid and liquid entropies are plotted in Fig. 19, where it can be seen that the two curves do indeed cross at exactly the temperature of the minimum in the melting curve. (The sudden descent toward zero of the solid entropy, in accordance with the third law, corresponds to an antiferromagnetic ordering transition at about 1 mK.)

The observation that the entropy of solid ^{3}He can sometimes be lower than that of the liquid appears at first sight to be quite astonishing considering that entropy gives a measure of the disorder in a system and that a liquid is usually considered to be in a highly disordered state compared to a crystalline solid. The resolution of this apparent paradox lies in the fact that, actually, liquid ^{3}He really is a highly ordered system but one in which the ordering takes place in momentum space, rather than in ordinary Cartesian space.

The existence of the minimum in the melting curve of ^{3}He is of considerable importance to cryogenic technology, in addition to its purely scientific interest, in that it provides the basis for the cooling method known as *Pomerancluk refrigeration*, a topic to which we will return.

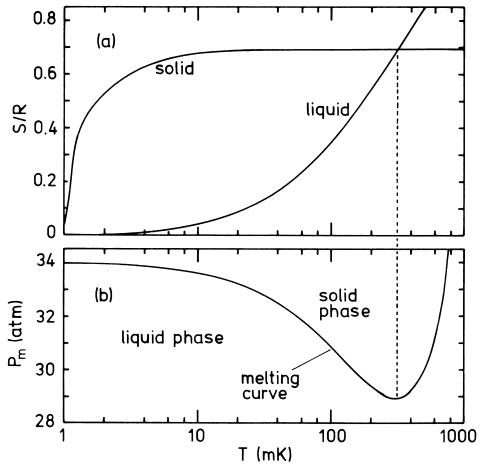


FIGURE 19 (a) The entropy S of liquid and solid ^{3}He (upper diagram) measured in units of the gas constant R, plotted as functions of temperature T and the melting pressure P_m of ^{3}He as a function of T. It may be noted that the minimum in the melting curve occurs at the temperature where the difference between the solid and liquid entropies changes sign. [From values tabulated by Betts, D. S. (1976). "Refrigeration and Thermometry Below One Kelvin," Sussex Univ. Press, London.]

D. Superfluid Properties of Liquid ^{3}He

Very soon after the publication of the BCS theory of superconductivity in 1957, it was realized that a similar phenomenon could perhaps occur in liquid ^{3}He. That is, the system might be able to reduce its energy if the ^{3}He atoms formed Cooper pairs. It was quickly appreciated, however, that the hard-core repulsion of the relatively large-diameter ^{3}He atoms would preclude the formation of pairs with zero relative orbital angular momentum ($L = 0$). Cooper paris with finite orbital angular momenta might, however, be energetically fovored but the temperature of the transition was difficult to predict with confidence. The phenomenon was eventually discovered at Cornell University in 1972 by Osheroff, who noted some unexpected anomalies in the otherwise smooth pressurization characteristics of a Pomeranchuk (compressional cooling) cryostat. Follwing that initial discovery, an immense amount of research has taken place and the superfluid phases of the liquid are now understood in remarkable detail. They share many features in common with He II but are vastly more complicted on account of their inherent anisotropy and their magnetic properties.

The low-temperature part of the ^{3}He phase diagram is shown in Fig. 20. It is very strongly influenced by the application of a magnetic field. In Fig. 20(a), with zero magnetic filed, there are two distinct superfluid phases

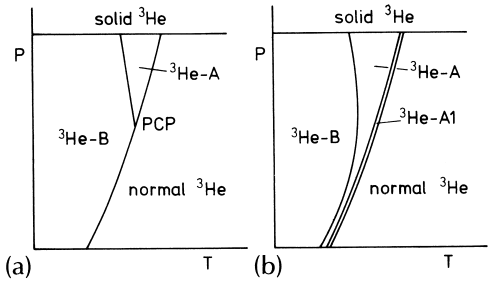


FIGURE 20 Schematic diagram illustrating the effect of a magnetic field on the low-temperature region of the ^{3}He phase diagram. In (a) there is no magnetic field and in (b) the magnetic field has been applied.

designated ^{3}He-A and ^{3}He-B, and there is a unique point, the polycritical point (PCP), where ^{3}He-A, ^{3}He-B and normal ^{3}He all co-exist in equilibrium. The application of a magnetic field causes this picture to change dramatically, as shown in Fig. 20(b). The PCP has then disappeared and, instead, a narrow slice of A-phase extends right down to meet the temperature axis at zero pressure. Simultaneously, a very narrow region of a third superfluid phase, the Al-phase, has appeared separating the A-phase from the normal Fermi liquid. As the magnetic field is further increased. the region of B-phase retreats towards lower temperatures until, at a field of $\sim 0.6T$, it has disappeared entirely.

The specific heat of liquid ^{3}He undergoes a finite discontinuity at the superfluid transition. This behavior is utterly different from the lambda transition (a logarithmic singularity) observed for liquid ^{3}He, shown in Fig. 10. It is, however, precisely the behavior expected at the transition to a BCS-like state; it is a beautiful example of a second-order phase transformation and is closely similar to the specific heat anomaly observed at the normal/superconducting transition for a metal in a zero magnetic field. When the liquid is cooled in a finite magnetic field, a pair of second-order transitions is seen, one for normal/Al and the other for the Al/A transition. The A/B transition is completely different in character, however, being a first-order phase transition (like boiling or freezing) in which a significant degree of supercooling can occur as the temperature of the liquid is being reduced.

Numerous experiments, and particularly those based on nuclear magnetic resonance (NMR) have shown that the superfluid phases of ^{3}He involve Cooper pairs in which the nuclear spins are parallel (equal spin pairing, or ESP, states with net $\mathbf{S} = 1$) and for which the orbital angular momentum quantum number $L = 1$. The Cooper pairs are thus rather like diatonic molecules, with atoms orbiting each other in states of finite relative angular momentum.

For Cooper pairs with $\mathbf{S} = 1$, $L = 1$ there are a very large number of possible states for the liquid as a whole. It has turned out that of the three possible spin projection values $\mathbf{S}_z = 0$ or ± 1, on the quantization axis, only $\mathbf{S}_z = \pm 1$ pairs are present in the A-phase; all three types of pairs are present in the B-phase; and only $\mathbf{S}_z = \pm 1$ pairs are present in the Al-phase. In each case, the relative orientations of the spin and orbital angular momenta adjust themselves so as to minimize the free energy of the liquid as a whole, and they do so in quite different ways for the A- and B-phases.

In the A-phase (and the Al-phase), the orbital angular momentum vector for every pair is orientated in the same direction, so the liquid as a whole must carry a resultant intrinsic angular momentum **I**. The A-phase is strongly anisotropic, with an energy gap $\Delta(T)$ that varies from zero in the direction of **I** to a maximum value in the direction of **S**, perpendicular to **I**. There is a superfluid order parameter **d**, analogous to the (scalar) macroscopic wave function in He II, that points in the same direction as **I**. The vector **I** intersects boundaries at right angles, but it also tends to orientate perpendicular to **S** and thus perpendicular to an external magnetic field. The liquid therefore acquires *textures* in much the same way as liquid crystals.

The B-phase has an isotropic energy gap, but the direction of **d** varies for different points on the Fermi surface. The liquid itself is not isotropic, however, and possesses a symmetry axis **N** that tends to align with external magnetic fields and is analogous to the directorix vector of a liquid crystal.

There is overwhelming evidence from numerous experiments for the superfluidity of both the A- and B-phases. One convincing demonstration is shown in Fig. 21, which plots the damping of a wire vibrating in the liquid as a function of temperature. Just above T_c, the viscosity of the liquid is comparable with that of light machine oil and rising so fast as T is reduced that the damping at T_c itself was too large to measure. Below T_c, however, the damping fell steadily with T to reach a value that was actually *smaller* than had been measured in a good vacuum at room temperature.

Most of the superfluid properties of He II (such as the creeping film, second-sound, and so on) have now also been demonstrated for superfluid ^{3}He. Like He II, the liquid can be described in terms of a two-fluid hydrodynamics. The ^{3}He superfluid component is associated with the condensate of Cooper pairs, and the normal fluid component with unpaired fermions in excited states. There is a Landau critical velocity representing the onset condition for pair-breaking, about 10^{-3} of the size of v_L for He II. The two-fluid properties of superfluid ^{3}He are greatly complicated by its anisotropy, particularly in the case of the

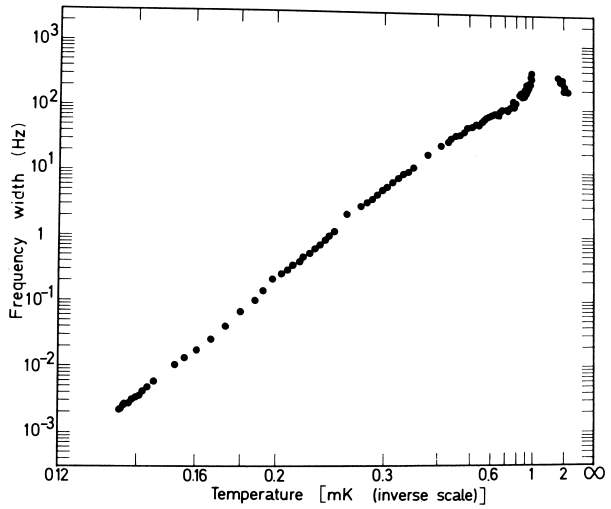


FIGURE 21 The damping experienced by a vibrating wire immersed in liquid ^{3}He under a pressure of 0.0 atm. The measured frequency width $\Delta\nu$ of the resonance, which gives a measure of the damping, fell by more than five orders of magnitude as the sample was cooled from T_c (1.04 mK) to the minimum temperature reached (0.14 mK). [From Guénault, A. M., Keith, V., Kennedy, C. J., Miller, I. E., and Pickett, G. R. (1983). *Nature* **302,** 695.]

A-phase where quantities such as the superfluid density vary with direction.

Superfluid ^{3}He has a wide range of interesting and important magnetic properties (which were of particular importance in the original identification of the physical nature of the phases). For example, both the A and B phases possess characteristic frequencies corresponding to oscillations of the **d** vector about its equilibrium orientation. They may be measured either by NMR, when the A-phase exhibits an extraordinary longitudinal resonance (with parallel static and radiofrequency fields), or by measurement of the frequency of parallel ringing. The latter phenomenon, known as the *Leggett effect*, is a very weakly damped magnetic oscillation of the liquid that occurs when the external magnetic field is stepped up or down by a small increment. It can be understood in terms of a transfer, by tunneling, of Cooper pairs between what are effectively separate but weakly coupled superfluids composed respectively of $\mathbf{S}_z = +1$ or $\mathbf{S}_z = -1$ pairs; that is, a kind of internal Josephson effect within the liquid.

IV. ACHIEVEMENT AND MEASUREMENT OF LOW TEMPERATURES

A. Evaporation Cryostats

Cryogenic temperatures are most easily achieved and maintained with the aid of liquefied gases, or *cryogens*, the most commonly used being nitrogen and helium. They may be liquefied on site or may be purchased as liquid. In either case, they are stored in highly insulated vessels prior to their use for cooling devices or samples of material in cryostats. For temperatures between the normal boiling points of N_2 and ^{4}He (77 K and 4.2 K, respectively), continuous-flow helium cryostats are often employed, with the temperature being adjusted by controlling the rate at which cold helium vapor removes the heat flowing in to the sample from room temperature. Table II summarizes the main cooling methods available in the temperature range below 4 K.

To reach temperatures below its normal boiling point, any given cryogen can be pumped so that it boils under reduced pressure. Nitrogen becomes solid at 63 K, however; its vapor pressure is so small that this temperature is the effective lower limit for a nitrogen cryostat. Liquid hydrogen solidifies at about 14 K, but the solid has a remarkably high vapor pressure and the temperature can be further reduced to about 10 K by sublimation under reduced pressure; on safety grounds, however, liquid hydrogen is seldom used in the laboratory. Liquid ^{4}He, as already mentioned, never solidifies under its own vapor pressure. The temperature reached is limited, instead, very largely by superfluid film flow; the creeping film tends to climb up the walls or up the pumping tube, evaporates in the warmer regions, and consequently saturates the vacuum pump. For this reason, the practical lower limit for a pumped ^{4}He cryostat is about 1 K, although with a diaphragm to reduce the perimeter over which the film flows and with the aid of very powerful pumps, lower temperatures may be achieved.

A typical liquid ^{4}He cryostat dewar is sketched in Fig. 22. Particular care is taken to minimize heat fluxes flowing into the liquid ^{4}He on account of its tiny latent heat; 1 watt is sufficient to evaporate 1 liter of liquid in approximately 1 hour. The inner vessel of liquid ^{4}He (and

TABLE II Some Cooling Techniques Available for Use Below 4 K

Technique	Temperature limits (K)	
	Upper	Lower
^{4}He evaporation[a]	4	1
^{3}He evaporation	1	3×10^{-1}
Dilution refrigeration	5×10^{-1}	3×10^{-3}
Pomeranchuk cooling	3×10^{-1}	10^{-3}
Adiabatic demagnetization[b] (electronic)	1	10^{-3}
Adiabatic demagnetization (nuclear)	10^{-2}	10^{-7}

[a] The lower limit can be pushed below 0.5 K by use of huge pumping speeds, but the rate at which heat can be extracted becomes too small to be useful.

[b] Limits shown are for commonly used salts.

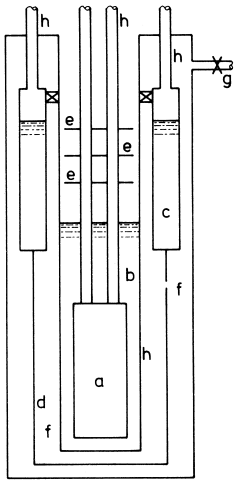


FIGURE 22 Schematic diagram of a typical metal helium cryostat dewar: a—insert with sample/device to be cooled; b—liquid ^{4}He at 4.2 K; c—liquid N_2 at 77 K in annular can; d—copper radiation shield at 77 K; e—gas-cooled copper radiation shields; f—vacuum; g—pump-out port for vacuum spaces; h—thin-walled stainless-steel tubes. The insert is usually sealed to the top of the dewar with a top plate and rubber O-ring (not illustrated).

most of the other vertical tubes, also) is made of thin-walled stainless steel in order to minimize the inflow of heat through conduction. For the same reason, it is surrounded by a vacuum space. There is then an annular liquid nitrogen vessel at 77 K, in order to intercept radiation from room temperature, and an outer vacuum space to protect the liquid nitrogen from thermal conduction from room temperature. Some radiation shields placed in the cold outgoing gas are used to prevent room-temperature radiation from shining directly on the cold space at the bottom. The whole helium bath may be pumped in order to adjust the temperature; more commonly, a dewar such as the one in Fig. 22 is used as the starting point for the other refrigeration methods described below. A ^{3}He evaporator or dilution refrigerator, for example, may be immersed in liquid helium in the dewar, boiling under atmospheric pressure at 4.2 K.

For temperatures between 0.3 K and 1.0 K it is convenient to use a ^{3}He evaporation cryostat. Liquid ^{3}He can be pumped to a lower temperature than ^{4}He, partly because of its larger vapor pressure at any given temperature, on account of its lower atomic mass and partly because of the absence of a creeping superfluid film. The material is expensive (at the time of writing, ~$350 for 1 liter of gas at STP, which yields ~1 cm^3 of liquid), and so is usually used in rather small quantities. A typical "single-shot" ^{3}He cryostat is sketched in Fig. 23. The ^{3}He gas, in total usually 1–5 liters at STP, passes down the ^{3}He pumping tube. It cools as it goes through the region where the tube is immersed in liquid helium at 4.2 K; it condenses in thermal contact with a pumped pot of liquid ^{4}He at ~1.1 K and then runs down into the ^{3}He pot. When condensation is complete, the ^{3}He pot is pumped. Its temperature can readily be reduced to ~0.3 K by use of an oil-diffusion pump. Depending on the influx of heat, the ^{3}He will typically last for several hours before recondensation is needed. Alternatively, the ^{3}He can be returned continuously to the pot from the outlet of the pump, via a capillary tube, with a restriction in the capillary to support the pressure difference of about 1 atmosphere between its top and bottom. In this mode, the cryostat has a higher base temperature which it can, however, maintain for an indefinite period.

B. Helium Dilution Refrigerators

Helium dilution refrigerators can provide continuous cooling within the remarkably large temperature range from about 0.5 K down to about 3 mK. The technique is based on the fact of there being a negative heat of solution when liquid ^{3}He dissolves in liquid ^{4}He. That the method can be carried to such extremely low temperatures is due to the fact that the limiting solubility of ^{3}He in ^{4}He, on the left-hand side of the phase separation diagram (Fig. 24), approaches 6% as $T \rightarrow 0$ rather than approaching zero (as happens for ^{4}He in ^{3}He on the right-hand side). There is no conflict with the third law, however, because the ^{3}He atoms have become highly ordered in momentum space, being a degenerate Fermi system with a well-defined Fermi sphere.

In fact, for concentrations of 6% and below, the properties of a ^{3}He–^{4}He solution at low temperatures are

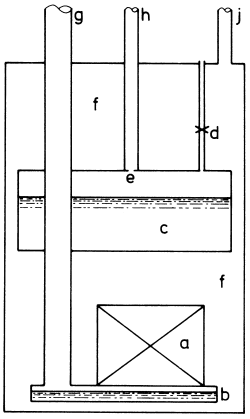


FIGURE 23 Schematic diagram of lower part of a typical single-shot ^{3}He cryostat: a—sample to be cooled; b—liquid ^{3}He at 0.3 K in copper pot; c—liquid ^{4}He at 1.1 K; d—needle-valve on stainless-steel filling tube for liquid ^{4}He; e—restriction to reduce superfluid film flow; f—vacuum; g, h, j—stainless-steel pumping tubes for ^{3}He, ^{4}He, and high vacuum, respectively.

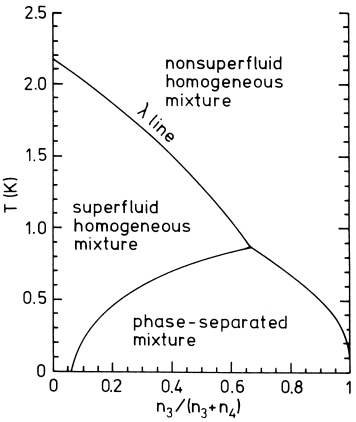


FIGURE 24 Phase-separation for liquid ^{3}He–^{4}He mixtures under their saturated vapor pressure, where n_3 and n_4 are the number densities of ^{3}He and ^{4}He, respectively, and T is the temperature. The finite solubility of ^{3}He in ^{4}He at $T = 0$ K should be particularly noted.

dominated by the ^{3}He. The liquid ^{4}He is almost 100% superfluid component through which the ^{3}He atoms can move freely, quite unhindered by any viscous effects. The ^{4}He can therefore be considered as a kind of background or ether whose only significant influences on the ^{3}He atoms are to increase their effective mass by a factor of about 2.5 and, of course, to prevent them from condensing. The properties of the ^{3}He are almost identical to those calculated for an ideal Fermi gas of the same density. When ^{3}He dissolves in pure liquid ^{4}He at a very low temperature it is as though it were "evaporating" in a "vacuum"; there is a corresponding latent heat, and it is this that is exploited in the dilution refrigerator.

Dilution refrigerators have been built successfully in a variety of different configurations, but the arrangement sketched in Fig. 25 is by far the most common one used in practice. In operation, the cooling effect occurs in the mixing chamber as ^{3}He atoms cross the interface between the concentrated ^{3}He-rich phase floating on top of the dilute ^{4}He-rich phase; the "evaporation" therefore takes place downwards. The dilute phase is connected to a still, which is usually held at about 0.7 K. When the still is pumped, the vapor that is removed is predominantly ^{3}He, because its vapor pressure is so much greater than that of ^{4}He. The depletion of ^{3}He within the still gives rise to an osmotic pressure, driving ^{3}He toward the still from the mixing chamber, where more ^{3}He can then dissolve across the phase boundary. The ^{3}He pumped off from the still is subsequently returned to the mixing chamber via a heat exchanger, so that it is cooled by the cold outgoing stream of ^{3}He on the dilute side. The art of designing efficient dilution refrigerators lies principally in the optimization of the heat exchanger.

The dilution refrigerator has become the work-horse of ultralow temperature physics. Its great virtues are, first, the huge temperature range that can be covered and, second, the fact that the cooling effect is continuous. Experiments can thus be kept at a temperature of a few mK for weeks or months at a time. In practice, the dilution refrigerator provides the usual starting point for additional "one shot" cooling by adiabatic nuclear demagnetization (see below), enabling experiments to be carried down to temperatures deep in the microkelvin range.

C. Pomeranchuk Refrigeration

Pomeranchuk cooling enables temperatures as low as 2 mK to be achieved and maintained, typically for a few hours. The technique exploits the negative latent heat of solidification of the ^{3}He that arises from the Clausius–Clapeyron equation and the fact that $(dP/dT)_m$ is negative for ^{3}He when $T < 0.3$ K (see above). If a volume of liquid ^{3}He is progressively compressed, therefore, it will cool as soon as solid starts to form, provided $T < 0.3$ K. When all of the liquid has been solidified, then the experiment is over and the cell must be allowed to warm

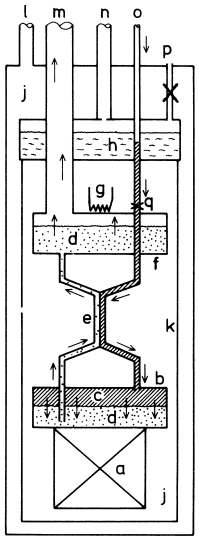


FIGURE 25 Schematic diagram illustrating the mode of operation of the standard dilution refrigerator. The concentrated (^{3}He-rich) phase is diagonally shaded and the dilute (^{4}He-rich) phase is dotted, the arrows indicating the direction of the ^{3}He circulation. a—sample to be cooled; b—mixing chamber between approximately 3 mK and 0.8 K; c—almost pure ^{3}He; d—dilute (approximately 6%) solution of ^{3}He in ^{4}He; e—heat exchanger; f—still at approximately 0.8 K; g—heater; h—pumped pot of liquid ^{4}He at approximately 1.2 K; j—vacuum; k—radiation shield at 1.2 K; 1, m, n—stainless-steel pumping tubes for high vaccum, ^{3}He and ^{4}He, respectively; o—stainless-steel condensing-in tube for ^{3}He returned from pump; p—stainless-steel filling tube for ^{4}He pot, with needle-valve; q—restriction.

up again as the solid decompresses before a new cycle can start.

The main problem in practice is to compress the liquid/solid mixture smoothly, without jerking, which would introduce heat. This can be accomplished through the use of liquid ^{4}He as a hydraulic medium to effect compression of the ^{3}He on the other side of a bellows. Because the solidification pressure of ^{4}He is considerably lower than that of ^{3}He (see Fig. 9), it is actually necessary to use, for example, two sets of bellows to form a hydraulic amplifier.

The great merit of Pomeranchuk refrigeration is the relatively large amount of heat that can be removed, being about ten times greater than in dilution refrigeration for any given ^{3}He circulation or solidification rate. It is particularly suitable for studies of ^{3}He itself, given the difficulties of heat transfer at very low temperatures and the fact that the cooling effect occurs within the sample under investigation. As already mentioned above, the original discovery of the superfluid phases of liquid ^{3}He took place in a Pomeranchuk cell. An inherent limitation of such studies, however, is that they are restricted to the melting pressure.

D. Magnetic Cooling

The technique of adiabatic demagnetization has been used for temperatures from 1 K down to the lowest attainable. Following the successful development of dilution refrigerators capable of providing useful continuous cooling powers below 10 mK the method is now used principally to cover the temperature range below ~5 mK. It is inherently a single-shot procedure. A variety of materials have been used as working media for demagnetization, but the basis of the technique is the same. The medium must be paramagnetic; that is, it must contain magnetic elements that can be orientated by an externally applied magnetic field thereby increasing the order and decreasing the entropy of the system.

The cooling procedure is sketched in Fig. 26, where it can be seen that, for any given temperature, the entropy is smaller when the magnetic field has been applied. The first stage is (ideally) an isothermal magnetization of the sample; the heat liberated as a result must be removed by, for example, the dilution refrigerator used for precooling the system. Once the magnetic field has fully been applied, the thermal link between it and the dilution refrigerator is broken, and the magnetic field is gradually reduced. Because the sample is now thermally isolated, this process is adiabatic and, hence, ideally, isentropic. The sample must therefore cool in order to get back to the B = 0 entropy curve again.

The temperature that can be reached must be limited ultimately by a ferromagnetic (or antiferromagnetic) ordering transition among the magnetic elements. It is not possible to cool the sample very far below the transition. Materials whose paramagnetism arises from electron spins, such as cerium magnesium nitrate (CMN), can be used to cool experiments to temperatures in the mK range, CMN itself being one of the best known and allowing temperatures near 1 mK to be attained. For lower temperatures than this, it is essential to choose a cryogenic medium whose magnetic elements interact less strongly with each other. The interactions between the magnetic Ce^{3+} ions in the CMN may, for example, be reduced by increasing their average separation. This can be accomplished by dilution of the CMN with the (nonmagnetic) lanthanum magnesium nitrate to produce a mixture known as CLMN. Although lower temperatures may thereby be attained, in principle, because of the resultant depression of the ordering temperature to well below 1 mK there is less available cooling power because of the reduced number of magnetic ions. A better approach is to use a nuclear paramagnetic system, such as the nuclei of copper atoms. The magnetic moments of the nuclei are so small that the interactions between them are negligible near 1 mK; the ordering transition takes place at ~100 nK. On the other hand, precisely because of the small nuclear moment, it is relatively difficult to align the nuclei with an external magnetic field. Successful nuclear adiabatic demagnetization therefore requires the largest attainable fields (in practice ~10 T) and the lowest possible starting temperatures (in practice ~7 mK, if the precooling time is not to be too long).

The main experimental problem in the μK and low mK ranges lies not so much in providing sufficient cooling power, for which the cryogenic technology is now readily available, as in promoting heat flow between the system under investigation and the refrigerant. For example,

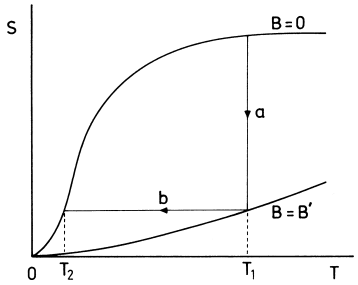


FIGURE 26 Schematic diagram illustrating the working principles of cooling by adiabatic demagnetization. The entropy S of the paramagnetic material is plotted against temperature T for zero applied magnetic field B, and for the large value $B = B'$. The system is precooled (e.g., by means of a dilution refrigerator) to $T = T_1$ and the external field is applied isothermally (a). The thermal link to the dilution refrigerator is then broken and the magnetic field is removed adiabatically, i.e., isentropically (b), when the system cools to the reduced final temperature T_2.

the nuclear spin system of copper is relatively easily and quickly cooled to 10^{-6} K but it is very much more difficult to cool even the conduction electrons in the same piece of copper and, more importantly, via the electrons and the lattice, to cool the item under study. This is because there is a thermal boundary resistance between all materials which becomes enormous at ultralow temperatures and implies the existence of a temperature discontinuity across any boundary across which heat is flowing. Thus, it is often the heat leak into the sample in the nuclear demagnetization cryostar that determines the ultimate temperature to which it can be cooled. By very careful vibration isolation, the screening of external electromagnetic fields, and judicious design and choice of materials, it is possible to reduce this heat leak to well below 1 nW. Such precautions have enabled liquid ^{3}He, for example, to be studied down to a temperature of approximately 120 μK (see Fig. 21).

An essential feature of any adiabatic demagnetization cryostat is a breakable thermal link between the refrigerant and the mixing chamber of the dilution refrigerator (or other precooler). A superconducting link of, for example, tin has been found particularly convenient and effective. It has a high thermal conductivity when driven into the normal state by an applied magnetic field but an extremely low thermal conductivity when in the superconducting state. The refrigerant can therefore be connected thermally to the mixing chamber or thermally isolated, simply by switching the magnetic field at the tin link on or off.

The most common nuclear refrigerant is copper, usually powdered or in the form of a bundle of wires, in order to reduce eddy current heating during demagnetization. Other nuclear refrigerants have also been used including. particularly, enhanced hyperfine interaction systems such as $PrNi_5$, which provides a much larger cooling power but cannot reach as low a final temperature as copper.

E. Thermometry

A very wide range of thermometric devices has been developed for the measurement of cryogenic temperatures. Thermometry between 4 K and room temperature is dealt with under *Cryogenic Process Engineering*; in this article, we discuss thermometry in the temperature range below 4 K. The principal thermometers and their approximate ranges of useful application are shown in Table III.

Vapor-pressure thermometry can only be based on ^{3}He or ^{4}He, being the only non-solid materials below 4 K. The pressure may be measured relatively easily by means of a manometer or McLeod gauge and then compared with international vapor pressure tables in order to find the temperature. The rapid variation of vapor pressure with temperature ensures that the latter can be determined to high precision from pressure measurements of modest accuracy.

TABLE III Some Cryogenic Thermometers Suitable for Use Below 4 K

Thermometer	Temperature limits (K)	
	Upper	Lower
^{4}He vapor pressure	4	1
^{3}He vapor pressure	1.5	5×10^{-1}
Carbon resistance	>4	10^{-3}
Germanium resistance	>4	5×10^{-2}
Thermocouples	>4	5×10^{-1}
Paramagnetic susceptibility[a] (electronic)	2	10^{-3}
Paramagnetic susceptibility (nuclear)	10^{-2}	$<10^{-6}$
Nuclear orientation	10^{-1}	10^{-3}
Noise thermometry	>4	10^{-2}

[a] For materials used in practice.

Resistance thermometry below 4 K is usually based on commercial carbon (radio) resistors or on germanium resistors. The measuring power must be kept small in order to prevent internal self-heating and varies in practice from about 10^{-7} W at 4 K down to 10^{-15} W in the low mK range. Both forms of resistor provide for rapid, convenient, and sensitive measurements, but both must be individually calibrated prior to use. Carbon resistors are usually recalibrated after each thermal cycle to room temperature but have the merits of cheapness and ready availability. Germanium resistors are much more stable over long periods and many thermal cycles, but are relatively expensive. In each case, more than one device is needed to cover the temperature range indicated with adequate sensitivity but without the resistance becoming too large for reliable measurement.

Thermocouples are generally best for temperature measurements above 4 K, where they are more sensitive, but some recently developed rhodium/iron alloys now permit measurements of useful sensitivity down to about 0.5 K. Once a given batch of material has been characterized, no further calibration is needed.

Paramagnetic susceptibility thermometers using salts such as CMN permit accurate temperature measurements down to about 1 mK and have the advantage that the technique becomes more sensitive as the temperature decreases. To a good approximation, over much of the range, a CMN thermometer follows Curie's Law and its susceptibility is inversely proportional to the absolute temperature; a single calibration point—for example, against ^{3}He vapor pressure at 0.6 K—is therefore in principle

sufficient to calibrate the device for measurements down to temperatures 50 times colder. At temperatures below 10 mK, however, due allowance must be made for the fact that Curie's Law is no longer accurately followed. The onset of the ordering transition provides a lower limit below which that particular susceptibility thermometer cannot be used. Thermometers based on the nuclear susceptibility of a metal can be employed to below 1 μK but require particularly sensitive measurement techniques because of the small magnitude of this quantity. Superconducting quantum interferometer devices (SQUIDs) can be employed to particular advantage in such measurements.

Nuclear orientation thermometry depends on the anisotropic emission of γ-rays from polarized radioactive nuclei. The extent of the anisotropy is detemined by the degree of polarization which, in turn, varies inversely with the absolute temperature. Since the nuclear hyperfine splittings are usually known to high precision, the thermometer is in principle absolute and does not require calibration. The useful temperature range is relatively narrow. however, being limited by radioactive self-heating at the lower end and insensitivity at the upper end.

Noise thermometry is another technique that, in principle, provides absolute measurements of temperature. The Johnson noise voltage V_{N} generated by the random (Brownian) motion of electrons in an element of resistance R is measured over a frequency bandwidth Δf. The mean-squared noise voltage $\langle V_n^2 \rangle = 4k_{\mathrm{B}}RT\Delta f$, from which T may be deduced if ΔF can also be measured. Unfortunately, a lack of sensitivity prevents the extension of this type of measurement below 10 mK where it would be most useful.

V. CRYOGENIC APPLICATIONS

Cryogenics is finding useful applications over an extraordinarily diverse range of engineering and technology. One of the most important and most widely exploited of all low-temperature phenomena is that of superconductivity, which is being applied to the construction of powerful magnets used for particle accelerators, for power storage, in medicine, and in superconducting motors and generators. It also holds out the promise of loss-free power transmission along superconducting cables. Superconducting instrumentation based on SQUIDs enables extraordinarily sensitive measurements to be made of very weak magnetic fields and is being widely used in archaeology, geology, and medicine.

The reduction of thermal (Johnson) noise in electronic circuits by cooling to cryogenic temperatures can yield extremely valuable improvements in signal/noise, thus facilitating the detection of very weak signals. Cooled infrared detectors are regularly used in astronomy, particularly for space and upper-atmosphere experiments.

The liquid heliums are being used to model phenomena that occurred in the very early universe, 10^{-35} s after the "big bang." There are theoretical reasons to expect that topological defects in space–time (e.g., cosmic strings) were formed as the universe fell through the critical temperature of 10^{27} K at which the GUT (grand unified theory) phase transition occurred. Cosmic strings are of particular interest and importance because, in some cosmologies, they are believed to have provided the primordial density inhomogeneities on which the galaxies later condensed. Unfortunately, defect formation at the GUT transition cannot be studied experimentally on account of the enormous energies involved, many orders of magnitude larger than available in even the most powerful particle acclerators, but the physics of the process can be studied by exploiting the close mathematical analogy that exists between the GUT and superfluid transitions. In practice, liquid helium is taken through the transition quickly and evidence is sought for the production of quantized vortices, which are the superfluid analogues of cosmic strings. At the time of writing, this process seems to have been reliably confirmed at the superfluid transition in ^{3}He but, surprisingly, not yet at the lambda transition in ^{4}He.

Isotopically pure liquid ^{4}He is being used as a down-scattering medium for the production and storage of ultracold neutrons (UCN)—neutrons of such low energy (and thus long de Broglie wavelength) that they can be totally reflected from certain materials. UCN can be held inside a neutrons, bottle and a range of experiments conducted on them. Because ^{3}He is a strong absorber of neutrons, the required isotopic ratio is extremely demanding, being $<10^{-12}$ compared with the $\sim 10^{-7}$ ratio typically found in natural well helium. Fortunately, purities of this order and better are readily achieved through use of superfluid heat flush to sweep away the ^{3}He atoms, which move with the normal fluid component.

Liquefied gases (O_2, N_2, H_2, He) find a wide range of applications in the steel industry, biology and medicine, the space program, and the food industry.

An interesting phenomenon known as the *quantum Hall effect*, which occurs in the two-dimensional electron gas in, for example, MOSFETs (metal-oxide-semiconductor-field-effect transistors), is becoming extremely valuable in metrology. At very low temperatures and strong magnetic fields, it is possible to exploit the effect to make absolute measurements of e^2/h (in effect, the atomic fine-structure constant) to very high precision.

SEE ALSO THE FOLLOWING ARTICLES

CHEMICAL THERMODYNAMICS • CRYOGENIC PROCESS ENGINEERING • MAGNETIC MATERIALS • QUANTUM THEORY • STATISTICAL MECHANICS • SUPERCONDUCTIVITY • THERMOMETRY

BIBLIOGRAPHY

Barone, A., and Paterno, G. (1982). "Physics and Applications of the Josephson Effect," Wiley, New York.

Barron, R. (1985). "Cryogenic Systems," 2nd ed. Oxford Univ. Press, London.

Betts, D. S. (1989). "An Introduction to Millikelvin Technology," Cambridge Univ. London.

Donnelly, R. J. (1990). "Quantized Vortices in He II," Cambridge University Press, London.

Hands, B. A., ed. (1986). "Cryogenic Engineering," Academic Press, London.

Lounasmaa, O. (1974). "Experimental Principles and Methods below 1 K," Academic Press, New York.

McClintock, P. V. E., Meredith, D. J., and Wigmore, J. K. (1984). "Matter at Low Temperatures," Blackie, Glasgow; Wiley-Interscience, New York.

Mendelssohn, K. (1977). "The Quest for Absolute Zero," 2nd ed. Taylor & Francis, London.

Shachtman, T. (1999). "Absolute Zero and the Conquest of Cold," Houghton Mifflin, Boston, MA.

Tilley, D. R., and Tilley, J. (1986). "Superfluidity and Superconductivity," 2nd ed. Hilger, London.

Vollhardt, D., and Wölfle, P. (1990). "The Superfluid Phases of Helium 3," Taylor & Francis, London.

Weisend, J. G. (1998). "Handbook of Cryogenic Engineering," Taylor & Francis, London.

White, G. K. (1979). "Experimental Techniques in Low Temparature Physics, " 3rd ed. Clarendon Press, Oxford.

Electrical and Electronic Measurements

Barrie A. Gregory, Retired
University of Brighton

GLOSSARY

Accuracy Quality that characterizes the ability of a measuring instrument to give indications equivalent to the true value of the quantity measured. The quantitative expression of accuracy is frequently given in terms of uncertainty.

Bandwidth Range of frequencies over which an instrument can be used. It is normally specified in terms of 3 dB points, that is, frequencies at which the response has fallen by 3 dB or 30% from the mid-frequency response.

Input impedance Impedance that a circuit under test "sees" when an instrument is connected to it. Input impedance is usually specified as a parallel resistance and capacitance combination, such as 1 MΩ shunted by 15 pF.

Insertion error Error introduced into a measured value as a result of the presence of an instrument.

Interference Signals that arise from sources extraneous to the measurement system and result in errors in the measured value.

Mean value Strictly, the rectified mean value, since this term is conventionally taken as the mean of the rectified value of an alternating waveform. It is 0.637 of the peak value of a sine wave.

Measured value Value indicated by an instrument or determined by a measurement process. It must be accompanied by a statement of the uncertainty or the possible limit of error associated with the measurement.

Measurement error Algebraic difference between the measured (or indicated) value and the true value.

Rise time Time interval measured from 10% to 90% of a step change.

Root mean square (rms) Effective value of an alternating current (or voltage), which is equivalent to the direct current (or voltage) value.

True value Actual or true value of a quantity cannot be determined, it can only be said to exist within the tolerance limits of a measured value.

Uncertainty Limits of tolerance placed on a measured value and obtained by adding the error contributions from all sources. All measured values should be accompanied by a tolerance or uncertainty figure.

ELECTRICAL AND ELECTRONIC measurement embraces all the devices and systems that utilize electrical and electronic phenomena in their operation to ascertain the dimensions, state, quality, or capacity of an unknown by comparison with a fixed known or reference quantity. The unknown or quantity being determined is commonly referred to as the *measurand* and the known or reference is termed the *standard*. The measurand or the standard may or may not be an electrical quantity, but it must be appreciated that electrical quantities are ultimately determined in terms of the primary standards of length, mass, and temperature.

I. METHODS AND TECHNIQUES

Within the definition of electrical measurement it has been stated that the magnitude of the measurand (or unknown) is ascertained by comparison with a reference quantity. It is therefore necessary in practice to have in every measuring instrument or system a known or reference quantity that may or may not have the same units as the measurand. For example, an unknown resistor may be compared with a known resistor, while a current may be compared with a force (e.g., a spring) or a voltage (*IR* drop). Hence, in measuring electrical quantities, various identifiable techniques have evolved as well as the need for reference, or standard, values and for a means of establishing how remote the measured value may be from the standard.

A. Standards and Absolute Measurements

So that instruments manufactured by companies in any part of the world will all provide the same value for a measured quantity it has been necessary to establish and maintain international and national standards to which the reliability of all measurements should be traceable.

The process of establishing international standards really began with the International Convention of the Metre in Paris in 1875. This convention, attended by the United States and most European countries, agreed on a particular platinum–iridium bar as the official meter and a cylinder of the same material as the kilogram standard for international use. These first international standards were created and kept at the International Bureau of Weights and Measures in Sèvres, France. In 1893 the United States decreed that its copies of these international standards would be their *fundamental standards* of length and mass. The turn of the century (1900) saw the establishment of the U.S. National Bureau of Standards (NBS) and the National Physical Laboratory (NPL) of the United Kingdom, which placed the responsibility for measurement standards in particular government departments. The establishment of

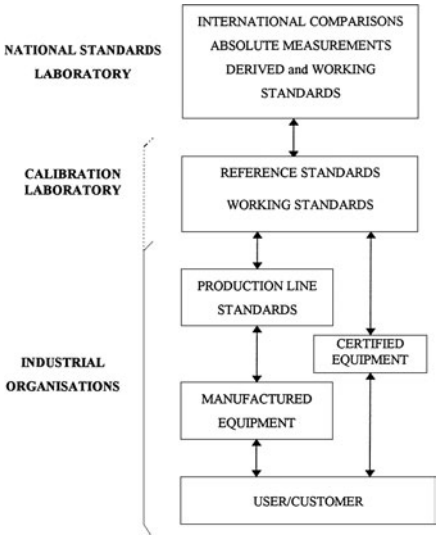


FIGURE 1 Traceability ladder showing the relationship between instruments in everyday use and the national standards.

standards laboratories has continued until, effectively, every industrialized country now has its own. While the prototype kilogram kept in Sèvres has remained the international standard of mass, the national standards laboratories in the various countries have, due to the requirement for ever more precisely defined standards, expended considerable effort to determine the fundamental quantities with as small an uncertainty as is possible within the limitations of technological development. Table I lists and defines the seven SI base units, and Table II shows the derived units used in electrical and electronic measurements.

Since it is not practical to check or calibrate the instruments in daily use against the national standards, traceability ladders of the type shown in Fig. 1 have evolved. Hence, national standards laboratories will have a set of "working standards" that are used to calibrate the reference or "transfer standards" of calibration laboratories. These, in turn, are used to check instrument manufacturers' working standards of, say, voltage, resistance, capacitance, and frequency, which are used for setting up the instruments sold to a customer. It is imperative that instrument users should operate a program of calibration for their equipment consisting of half-yearly or more frequent checks on all their instruments. Such a program has largely been ignored because of its expensive use of highly skilled personnel in lengthy, labor-intensive processes. However, the development of programmable calibration systems has appreciably reduced the time and operator skill required, thus making regular calibration checks feasible. Keeping calibration records is an essential part of the checking process, for it is from a study of the records that the likely errors in an instrument's readings can be established and that confidence in its performance can develop.

TABLE I The Seven SI (Système International) Base Units[a]

Quantity	Unit name	Unit symbol	Definition
Length	Meter	m	The length of the path traveled in vacuum by light during 1/299,792,458 of a second
Mass	Kilogram	kg	The mass of the international prototype
Time	Second	s (or sec)[b]	The duration of 9,192,631,770 periods of the radiation corresponding to the transition between the two hyperfine levels of the ground state of the cesium 133 atom
Electric current	Ampere	A	The constant current maintained in two parallel conductors of infinite length, negligible circular cross section, and 1 m apart in vacuum that produces between these conductors a force equal to 2×10^{-7} N per m of length
Thermodynamic temperature	Kelvin	K	The fraction 1/273.16 of the thermodynamic temperature of the triple point of water
Luminous intensity	Candela	cd	The luminous intensity, in a given direction, of a source which emits monochromatic radiation with a frequency 540×10^{12} Hz and whose energy intensity in that direction is 1/683 W/steradian
Quantity of substance	Mole	mol	The amount of substance of a system that contains as many specified elementary entities (electrons, ions, atoms) as there are atoms in 0.012 kg of carbon 12

[a] After IEE. (1997). "Units and Symbols for Electrical and Electronic Engineering."
[b] Where the use of s may cause confusion the abbreviation sec is recommended.

B. Measurement Methods

While all measurements require comparison of the measurand with a reference quantity, the method or process by which the comparison is made will vary according to the parameter under observation, its magnitude, and the conditions prevalent at the time of observation. The methods in use in electrical and electronic measurement can be categorized as substitution, force interaction, or derivation.

1. Substitution

As implied, this method requires an unknown to be replaced by a known value of such magnitude that conditions are restored to a reference condition. For example, consider the situation illustrated by Fig. 2. A voltage source causes a current I to flow through a circuit consisting of an ammeter in series with an unknown resistance R_x. If the switch is changed so that the unknown is replaced by a

TABLE II Derived Units Encountered in Electrical and Electronic Measurements[a]

Quantity	Unit name	Unit symbol	Derivation in term of SI base units
Force	Newton	N	$m\ kg\ s^{-2}$
Energy	Joule	J	$m^2\ kg\ s^{-2}$
Power	Watt	W	$m^2\ kg\ s^{-3}$
Pressure	Pascal	Pa	$m^{-1}kg\ s^{-2}$
Electric potential	Volt	V	$m^2\ kg\ s^{-3}\ A^{-1}$
Electric charge	Coulomb	C	sA
Magnetic flux	Weber	Wb	$m^2\ kg\ s^{-2}\ A^{-1}$
Magnetic flux density	Tesla	T	$kg\ s^{-2}\ A^{-1}$
Electric resistance	Ohm	Ω	$m^2kg\ s^{-3}\ A^{-2}$
Electric conductance	Siemens	S	$m^{-2}\ kg^{-1}\ s^3\ A^2$
Capacitance	Farad	F	$m^{-2}\ kg^{-1}\ s^4\ A^2$
Inductance	Henry	H	$m^2\ kg\ s^{-2}\ A^{-2}$
Frequency	Hertz	Hz	s^{-1} or sec^{-1}
Torque	Newton meter	Nm	$m^2\ kg\ s^{-2}$
Electric field strength	Volt per meter	V/m	$m\ kg\ s^{-3}\ A^{-1}$
Magnetic field strength	Ampere per meter	A/m	$m^{-1}\ A$
Thermal conductivity	Watt per kelvin	W/mK	$m\ kg\ s^{-3}\ K^{-1}$

[a] After IEE. (1997). "Units and Symbols for Electrical and Electronic Engineering."

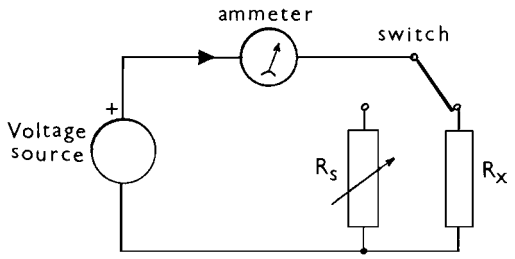


FIGURE 2 Possible circuit for the substitution method of resistance measurement.

decade resistance box R_S, the magnitudes of the decades can be adjusted until the current is restored to the value that was present when R_x was in the circuit. The R_S setting then becomes equal to the value of R_x. This process is only occasionally a satisfactory method of measuring a resistor, but in the Q meter (Section II.E) a substitution process is used for the measurement of small capacitance values, reference conditions being restored by reducing a known variable capacitor by an amount equal to the unknown capacitor.

2. Force Interaction

This method, which has been exploited since the development of the galvanometer in the early 1820s, is based on the interaction of the magnetic field surrounding a current-carrying conductor with another magnetic field to produce a deflecting force. By arranging the current-carrying conductor as a coil, constraining its movement to that of rotation within the field of a permanent magnet, and applying a control or opposing torque from a spiral spring, we obtain the moving coil that forms the active element of most analog instruments (Fig. 3). Other current-sensing instruments obtain the deflecting torque by the interaction of ferromagnetic vanes in a coil (*moving-iron instrument*) and through the interaction of the magnetic fields produced by a number of coils (*electrodynamometer instrument*). Disadvantages of both moving-iron and electrodynamometer instruments are that they have nonlinear scales and consume enough power in many situations to affect the performance of the circuit under test. An alternative to the electromechanical interaction is to derive the displayed deflection from the interaction of electrostatic forces between fixed and moving vanes; these latter instruments are now only used in high-voltage applications.

An advantage of this method of measurement is that the magnitude of the unknown signal is continuously monitored.

3. Derivation

This category may be subdivided into indirect methods, ratio techniques, and bridge or null methods. An example of the first is the use of Ohm's law to determine the value of a resistance from voltage and current measurement or the evaluation of a current from the voltage drop across a resistor, both of these techniques being commonly used in digital multimeters (see Sections II.D and II.B).

Ratio techniques have an advantage over the above in being inherently more accurate due to their comparison of like with like. For example, an unknown resistance R_x may be determined by comparing the voltage drop V_x across it with the drop V_S across a known resistance R_S when R_x and R_S are connected in series to a constant supply. Then $V_x = IR_x$ and $V_S = IR_S$, where I is the circuit current, and from these equations $R_x = R_S V_S / V_x$. If V_x and V_S are measured using the same voltmeter, the error in measurement is largely that associated with the reference resistor R_S since most of the error in the voltage measurement will cancel out. To illustrate this, assume V_S and V_x are both measured using a voltmeter that indicates δ% high; then

$$R_x = \frac{R_S(1+\delta)V_{x\,\text{true}}}{(1+\delta)V_{S\,\text{true}}} = \frac{R_S V_{x\,\text{true}}}{V_{S\,\text{true}}}$$

which shows that the need to know the absolute values of V_x and V_S has been removed. This is an important principle and is utilized in many modern electronic instruments.

A progression from the ratio method of measurement is the creation of a circuit so that the difference between the unknown and a proportion of the known is detected and then reduced to zero by a balancing or "null detection" process. An illustration of this technique is the operation of the dc potentiometer, which in its simplest form (Fig. 4), consists of a voltage source E driving a current I around the circuit $ABCD$. If BC is a 1 m length of wire having a resistance of 20 Ω and $I = 50$ mA, the voltage drop along BC is 1 V or 0.001 V/mm. Thus any direct voltage less than 1 V can be determined by adjusting the position of the slider to give zero deflection on the galvanometer G and measuring the length of wire between the wiper and the end C. While developments in digital electronic voltmeters have meant that the potentiometer has been superseded as a means of voltage measurement, the principle is still employed in a number of instruments, for example, in potentiometric pen recorders.

By combining the ratio and potentiometric principles, we can devise a circuit in which the voltage drops across known and unknown components (resistors) can be compared and made equal. Such circuits (Fig. 5) are termed *bridges* and are investigated further in Section II.D.

C. Analog Techniques

Analog measurements are those involved in continuously monitoring the magnitude of a signal or measurand. The use of analog instrumentation is extensive, and while

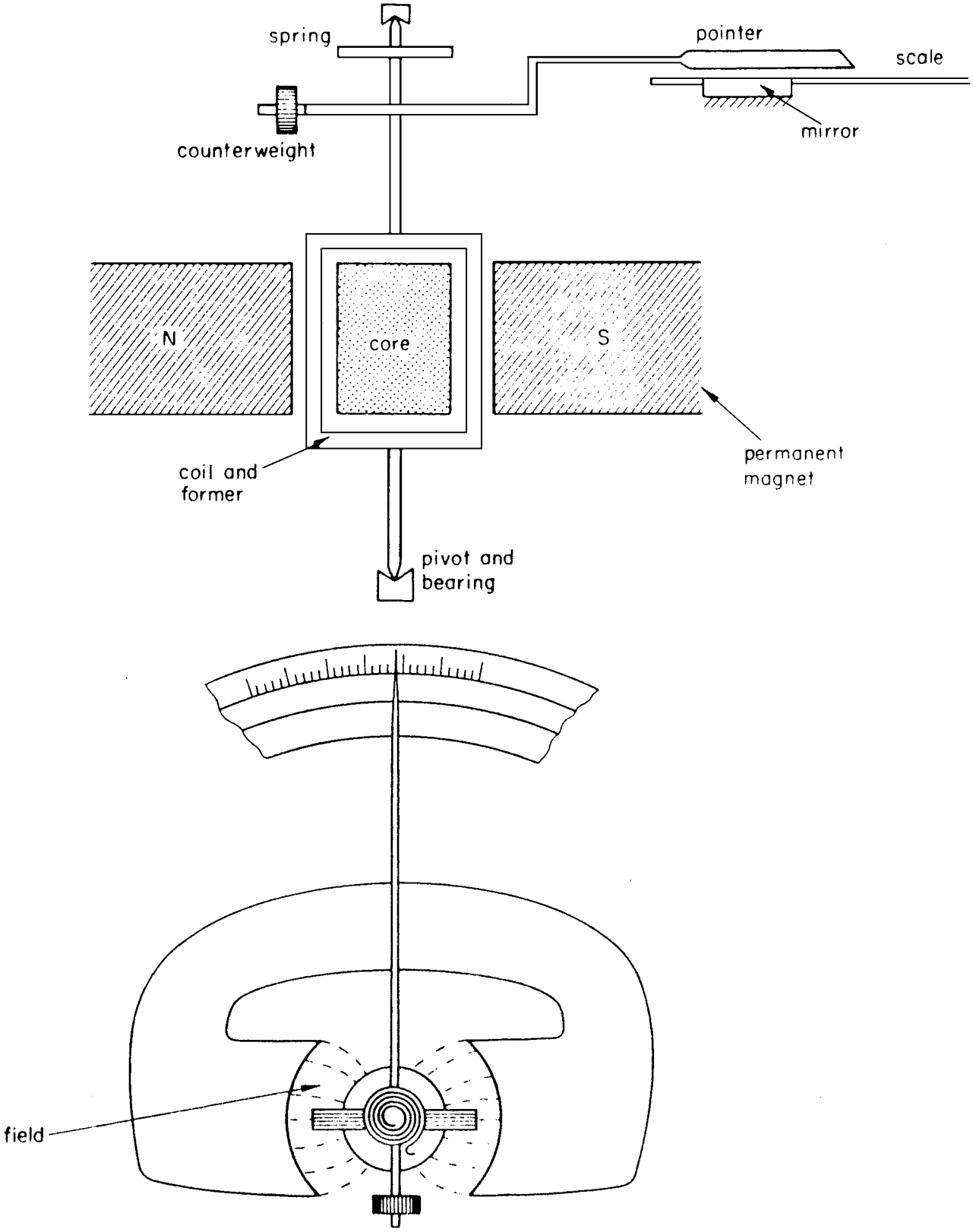


FIGURE 3 Moving-coil meter movement. [From Gregory, B. A. (1981). "An Introduction to Electrical Instrumentation and Measurement Systems," Macmillan Education, Hampshire, England.]

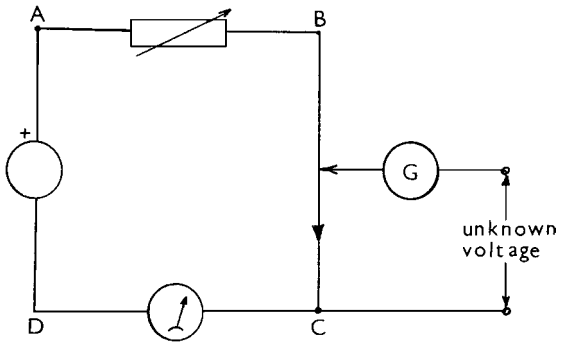


FIGURE 4 Elementary dc potentiometer circuit.

digital instruments are ever increasing in number, versatility, and application, it is likely that some analog devices will never be replaced by digital devices. The reason for this is that an operator can assimilate the information more rapidly from a multi-analog display than from a multi-digital display. In addition, it is very much easier to adjust a quantity to a particular value using an analog display as opposed to a digital one. However, a gradual increase seems very likely in the number of hybrid instruments, that is, devices that employ digital processes in their operation but have an analog display. A number of general-purpose multimeters currently produced have both a digital and a bargraph analog display as in Fig. 6 (see also Fig. 49).

A large number of the analog instruments in use are electromechanical in nature, making use of the force-interaction processes described in the previous section. The displays used in analog instruments involve either pointers and scales or some form of graphics. Considering first the pointer displays, we show the forms of scale in common use in Fig. 7. The instruments with a linear scale (equally spaced divisions) are generally the most convenient to use, particularly those with subdivisions in tenths. The *effective range* of an analog instrument is defined by that part of the scale over which readings can be made with the stated accuracy and should be recognizable without ambiguity. How this is accomplished is left to the manufacturer; Figs. 7c and 7d show two possible methods. One of the problems associated with pointer instruments is misreading due to parallax; that is, if the user

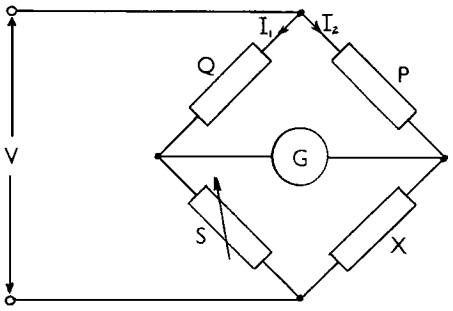


FIGURE 5 Basic arrangement of the Wheatstone bridge circuit.

FIGURE 6 Hand-held digital multimeter with bargraph display. (Courtesy of Schmidt Scientific Taiwan Ltd.)

is incorrectly positioned relative to the scale, an erroneous reading is made. To assist the user in removing this source of reading error, most quality instruments have a mirror adjacent to the scale so that the fine blade of the pointer can be aligned with its reflection. It should also be realized that the limit of resolution on any analog scale is approximately 0.3 mm, and readings that imply greater resolution are due to the operator's imagination.

The displays of graphical instruments are required to record the variation of one quantity with respect to another. The most widely used axes are function and time, usually referred to as y–t. Examples are the display of a voltage waveform on an oscilloscope (Section III.A) or the recording of temperature variations in a manufacturing process (Section III.B). Some instruments are designed so that the variations of one quantity with respect to another quantity may be recorded and these are referred to as x–y plotters. One particularly useful form of display in waveform analysis is the presentation of the amplitude of the frequency components of a signal versus frequency, an arrangement that is illustrated by the spectrum analyser (see Fig. 55).

If a graphical display is required as a function of a measurement system, a factor that must be considered is whether the record can be temporary or if it must be permanent. If the latter is required, either photographic techniques or some form of hard-copy writing system must be used. The filing of quantities of recorded data can create reference and storage problems, hence the use of temporary records is often desirable. Examples of this are virtual instruments (Fig. 11) and the storage oscilloscope (Section III.B).

D. Digital Measurements

Most digital instruments display the measurand in discrete numerals, thereby eliminating the parallax error and

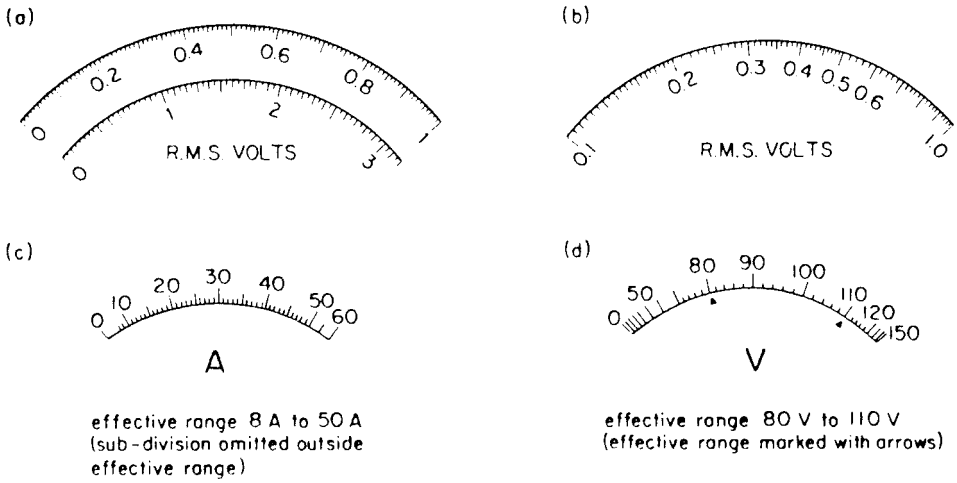


FIGURE 7 Examples of scales used on analog pointer instruments. [From Gregory, B. A. (1981). "An Introduction to Electrical Instrumentation and Measurement Systems," Macmillan Education, Hampshire, England.]

reducing the operator errors associated with analog pointer instruments. In general, digital instruments are more accurate than pointer instruments, and many (as in Fig. 6) incorporate function, automatic polarity, and range indication, which reduces operator training, measurement error, and possible instrument damage through overload. In addition to these features, many digital instruments have an output facility enabling permanent records of measurements to be made automatically.

Digital instruments are, however, sampling devices; that is, the displayed quantity is a discrete measurement made either at one instant in time or over an interval of time by digital electronic techniques. In using a sampling process to measure a continuous signal, care must be exercised to ensure that the sampling rate is sufficient to allow for all the variations in the measurand. If the sampling rate is too low, details of the fluctuations in the measurand will be lost, whereas if the sampling rate is too high, an unnecessary amount of data will be collected and processed. The limiting requirement for the frequency of sampling is that it must be at least twice the highest frequency component of the signal being sampled, so that the measurand can be reconstructed in its original form.

In many situations the amplitude of the measurand can be considered constant (e.g., the magnitude of a direct voltage). Then the sampling rate can be slowed down to one that simply confirms that the measurand does have a constant value. To convert an analog measurand to a digital display requires a process that converts the unknown to a direct voltage (i.e., zero frequency) and then to a digital signal, the exception to this being in the determination of time-dependent quantities (Section III.C).

The underlying principle of digital electronics requires the presence or absence of an electrical voltage; hence a number of conversion techniques are dependent on the counting of pulses over a derived period of time. Probably the most widely used conversion technique in electronic measurements incorporates this counting process in the *dual-slope* or *integrating* method of analog-to-digital (A–D) conversion. This process operates by sampling the direct voltage signal for a fixed period of time (usually for the duration of one cycle of supply frequency, i.e., 16.7 msec in the United States, 20 msec in the United Kingdom). During the sampling time, the unknown voltage is applied to an integrating circuit, causing a stored charge to increase at a rate proportional to the applied voltage, that is, a large applied voltage results in a steep slope, and a small applied voltage in a gentle slope (Fig. 8). At the end of the sample period, the unknown voltage is removed and replaced by a fixed or reference voltage of the opposite polarity to the measurand. This results in a

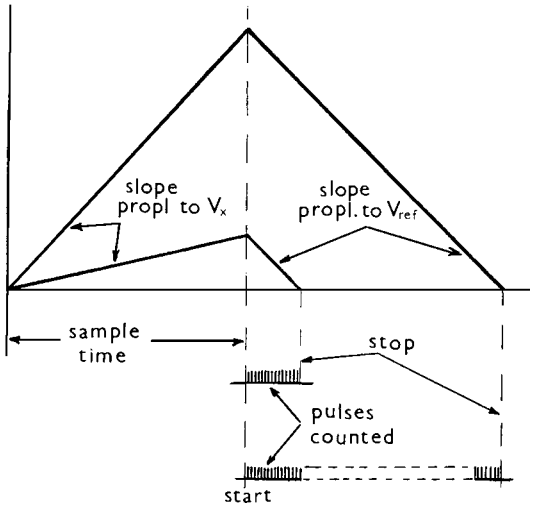


FIGURE 8 Dual-slope A–D conversion process.

reduction of the charge on the integrating circuit until zero charge is detected. During this discharge period, pulses are routed from a generator to a count circuit, which in turn is connected to a display. The reason for the wide use of this technique in digital multimeters is that, since the sampling time is one cycle of power supply frequency, any electrical interference of that frequency superimposed on the measurand will be averaged to zero.

Another A–D conversion technique commonly used in electronic instruments is based on the successive-approximation process. In this a fixed, or reference, voltage is compared with the unknown. If the unknown is greater than the reference (which could have a magnitude of, say, 16 units), this fact is recorded. The reference is then halved in value and added to the original reference magnitude, so that a known value of 24 units is compared with the measurand. Had the unknown been less than 16 units, a comparison between 8 units and the measurand would be made. Fractions of the reference level are added (or not) until an approximation to the magnitude of the measurand is established (within the limits of the available resolution). For the unknown voltage in Fig. 9, the magnitude is $16 + 8 + 0 + 0 + 1$ units or, if represented by a binary chain, 1101 units. The advantage of the successive-approximation technique is that it can be manufactured to have a very short sampling time (less than 1 μsec) and can thus be used for fast analog-to-digital conversions in, for example, waveform recording and analysis.

The principal advantage of a digital display is that it removes ambiguity, thereby eliminating a considerable amount of operator error or reading misinterpretation. Unfortunately, it may create a false confidence. The assumption that "I can see it, therefore it must be correct" may not be true due to interference effects, the specified limitations, and the loading effects of the instrument. The modern trend of incorporating programming capability

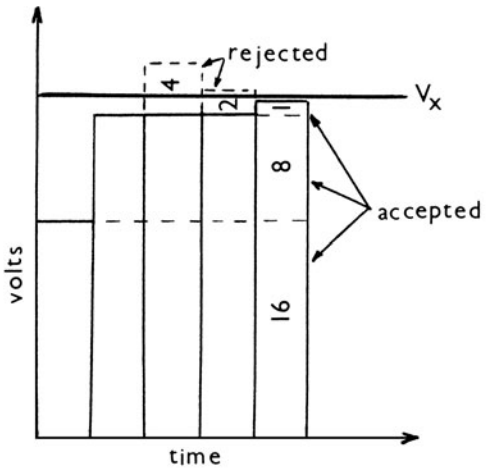


FIGURE 9 Successive approximation A–D conversion process.

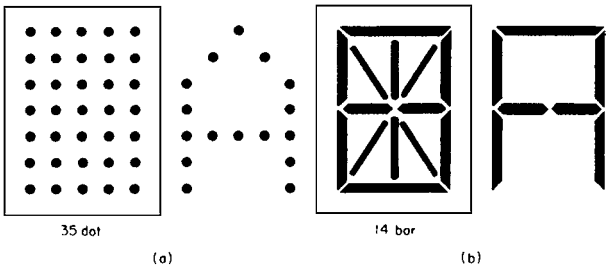


FIGURE 10 Digital displays: (a) dot matrix and (b) bar arrangement.

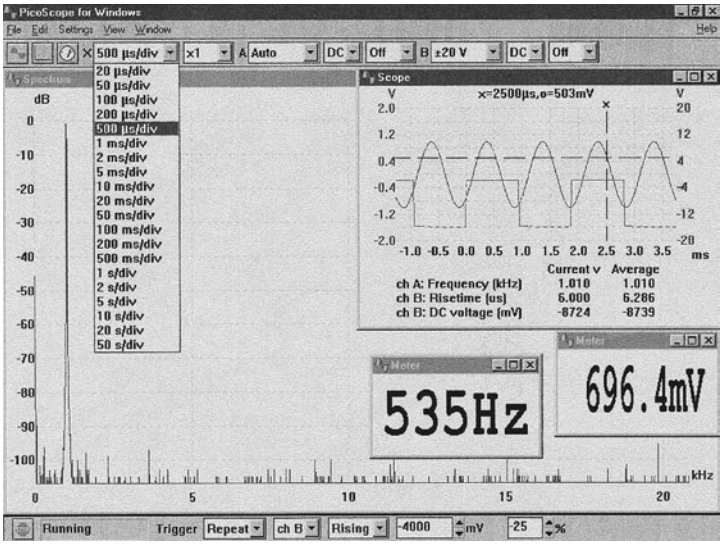
into instruments has created a need for alphanumeric displays rather than a simple numeric display. Figure 10 illustrates two types of array that are in use in addition to the seven-segment numeric display. The methods used to produce the display are light-emitting diodes (LEDs), liquid-crystal displays (LCDS), gas discharges (Nixie tubes and plasma panels), and cathode ray tubes (CRTs) or, as they are becoming more frequently known, visual display units (VDUs).

The developments in computing technology have led to the creation of "virtual instruments" where an interface to a laptop or personal computer is used to connect points in a circuit or system and the screen display contains several conventional instrument displays (Fig. 11). These virtual instrument systems vary in complexity from a terminal "pod" to which one or two points in a circuit may be connected and can provide a VDU display of signal magnitude, frequency, spectral analysis, and waveform, to more complex arrangements with plug-in cards that can take 8, 16, or 32 inputs and perform analysis and computations on the signals connected to the plug-in card.

E. External Influences

In many practical situations the conduction of a small electrical signal from the measurand to the measuring instrument will be affected by a number of forms of interference. These may be divided into (1) the effects of the environment on the component parts of the instrument or measuring system and (2) the injection of unwanted signals from unrelated electrical circuits and fields into the measuring system.

Thermal effects are the most common of the environmental effects, since all components will have a temperature coefficient of expansion and many will have a temperature coefficient of resistance. Thus, besides possible resistance changes with temperature, there will be dimension changes, with the resulting possibility of stress forces being applied to mounted components. Since semiconductor devices are pressure sensitive, ambient temperature changes may have a marked effect on the performance of equipment and circuits incorporating any such



(a)

(b)

FIGURE 11 Virtual instrument: (a) Display of waveforms, values, and spectrum; (b) terminal "pod." (Courtesy of Pico Technology Ltd., Cambridge, UK.)

devices. The normal variations in atmospheric pressure are generally of little consequence and can usually be neglected. Exceptions are in high-voltage measurements, in situations where air is used as a dielectric, and possible effects on electronic components used in aircraft. The humidity of the environment can also be of importance since some materials are hydroscopic and changes in humidity will change their electrical properties.

The major problems of injected signals are associated with alternating voltages and fields. However, two sources of direct voltage interference exist. First is that due to thermal emfs: voltages generated when two junctions of dissimilar materials are at different temperatures. While this phenomenon is usefully exploited for temperature measurement (Section IV.A), it can be a problem when measuring small signals, for voltages of a few microvolts will be generated for each degree Celsius of temperature difference between dissimilar material junctions. The second source of direct voltage interference results from the manufacturing processes of electronic equipment, it being normal to solder components onto copper strips on an insulating board. The presence of copper, solder, and

unremoved flux residuals will, if made damp by, say, condensation, provide a galvanic voltage that may be tens of millivolts: a severe problem!

Alternating interference signals result from the following:

1. *Electrostatic* (capacitive) coupling that exists between measuring circuits and devices that are at a different potential. This interference can be removed by the use of grounded conducting screens, hence the use of coaxial cable. *NB*: The screen must be at earth potential for this to be effective.
2. *Electromagnetic* (inductive) coupling resulting from the linking of the measuring circuits with the fields surrounding current-carrying conductors. This form of interference can be reduced by the use of magnetic shielding and careful component layout.
3. *Inherent noise* within the measuring circuits and instruments that occurs as a result of the random movement of electrons in conductors and components. This is a major problem only for the measurement of very small signals and is reduced by the use of (expensive) selected low-noise components.
4. *The mains or power supply*, that is, surges and spikes transmitted from other equipment and received by the measuring system via the power supply lines. This interference can be removed by the use of mains input filters.
5. *Radiated interference*, that is, radiofrequency signals transmitted by various sources, such as electrical discharges, CB radio, and authorized transmissions, that are picked up by the measuring system. These effects are removed by the use of conducting shields around sensitive parts of the measurement system.

F. Measurement Errors

No measurement is perfect. Even the most precise measurements conducted at a national standards laboratory do not exactly determine the magnitude of a quantity. An uncertainty or tolerance always exists on a measurement, and the magnitude of the uncertainty indicates the quality of measurement. Since the cost of a measurement may be considered to be determined by the inverse of the measurement tolerance, it is important that a measurement process appropriate to the application is used. For example, why use a measurement process costing $100 and giving a tolerance of $\pm 0.001\%$ when one costing $1 and yielding a tolerance of $\pm 1\%$ would be adequate.[1]

[1]In the United Kingdom there is a tendency to refer to the tolerance on an instrument's reading as its accuracy, e.g., $\pm 0.5\%$, whereas in the United States the accuracy *of the same instrument* is likely to be stated as 99.5%.

Thus, when any measurement is made, the first consideration should be: How precisely do I need to know the value? That is to say, what is an acceptable tolerance on the measurement? When the acceptable limit of tolerance on a measurement is established it is then necessary to identify all the possible sources of error in the instruments being used and the effects of the measurement system on the measurand. Such error sources are as follows.

1. Construction Errors

The reading on an instrument will be affected by imperfections in components in the instrument. If it is an electromechanical device, these errors will be due to magnetic hysteresis, friction, and tolerances on the sizes, assembly, and purity of the components. Likewise, for an electronic instrument, tolerances on components, assembly, and hysteresis of operation of the various circuits will affect the operation. In both types of instrument, any changes in the environment (temperature, humidity, and possibly pressure) will have an effect on the performance. Since many materials change their properties slightly with age (and continual use), it is necessary to consider the effect of age on the performance of an instrument. Since this is difficult to predict, it is essential that instruments be checked (calibrated) at regular intervals, for example, once a year, but preferably every 6 months. From the records (history) of instruments, confidence in the performance of a particular instrument is maintained.

2. Determination Error

This error is the uncertainty in the indicated value due to the resolution of the instrument. Determination error is dependent on the display method and thus will have as a minimum one of the following values:

(a) ± 0.3 mm on an analog scale or trace
(b) ± 1 count or least significant digit (*lsd*) in a digital display
(c) $\pm$half a unit of the least decade of a bridge arm (or decade box), assuming that a detector of sufficient sensitivity is in use.

Construction and determination errors are accounted for within the accuracy specification of an instrument: for example, $\pm 1\%$ of the full-scale deflection (FSD) for an analog instrument and either $\pm$(0.01% of reading $+ 1$ count) or $\pm$(0.01% of reading $+0.05\%$ of range) for a digital instrument.

Note that (a) all these tolerances should be for a specified temperature band (say $23 \pm 5^{\circ}$C), and when the instrument is used outside this band, an additional specified

tolerance per Celsius degree should be added; and (b) it is important to look carefully at the specification of an instrument to establish how its accuracy has been specified and apply the appropriate tolerance to all measurements made using it.

3. Calculation Errors

When a derivation process is used to evaluate the result of a measurement, calculations are inevitable. Since computation may at some stage involve truncation and/or rounding, these numerical processes may provide an error in the result. This form of error is usually of significance only if precise measurements are being made, for example, in time-dependent measurements in which tolerances of $\pm 0.001\%$ (or smaller) are common. Another source of calculation error results from simplification of formulas. An example of this is the balance equation for a Kelvin double bridge (Fig. 31) used for the measurement of low resistance which is simplified from

$$R_x = \frac{QS}{M} + \left(\frac{mr}{m+q+r}\right)\left(\frac{Q}{M} - m\right)$$

to

$$R_x = \frac{QS}{M}$$

on the assumption that $M = m$, $Q = q$, and $r = 0$.

4. Operation Errors

The sources of error in operating a measuring system result from insertion effects, system errors, and external influences. When an instrument is inserted into a circuit or system to perform a measurement, it changes the system. Hence there is a need to use voltmeters with as high an input resistance as possible and ammeters with as low an input resistance as possible. This condition is exacerbated by the effects of frequency on the input impedance of ac instruments that have a capacitive component and hence a low impedance at high frequencies ($Z = 1/\omega C$).

In assembling a measurement system, care must be exercised not to introduce factors that would result in operation errors. Examples of such error sources are lead and contact resistances (when measuring low resistances and small signals), uneven cable lengths when measuring small time differences, long connecting wires if operating at high frequencies, and dirty or damp surfaces when measuring high resistances.

The effects of external influences (interference) on the operation of a measurement system can be considerable and care must be taken to use suitable screened connections and instrument input types. If random external events are the possible cause of a problem, it may be necessary statistically to eliminate a rogue reading.

5. Loading or Insertion Error

Any instrument when connected into a circuit will change the conditions from those that existed in the circuit prior to its insertion. It is therefore important to ensure that this disturbance is made as small as possible, otherwise incorrect readings will be produced, that is, an error in addition to those inherent in the instrument will be added to the measurement.

Consider the simple circuit in Fig. 12 (which may be the equivalent circuit of a more complex arrangement). Let V_m be the voltage indicated by the meter, which has a resistance R_m, and is connected across a source V_s of internal resistance R_s. Then

$$V_m = V_s \frac{R_m}{R_m + R_s}$$

and the error in the reading resulting from the voltage division effect will be

$$\text{Insertion error} = \frac{V_m - V_s}{V_s} \times 100\%$$
$$= \frac{V_s[R_m/(R_m + R_s)] - V_s}{V_s} \times 100\%.$$

Hence

$$\text{Insertion error} = \frac{-R_s}{(R_m + R_s)} \times 100\%.$$

By making R_m equal to nR_s, Table III has been created, from which it can be seen that R_m needs to be very much greater than the source resistance for the insertion error to be negligible. This important concept needs to be borne in mind when, for example, a digital voltmeter that has an input resistance of 10 MΩ and a specified accuracy of $\pm 0.01\%$ is being used to measure the voltage at a test point in a circuit that is equivalent to a source that has an output resistance of 10 kΩ. From the table the insertion error is

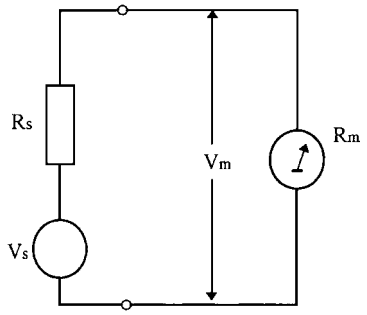


FIGURE 12 Equivalent circuit for the evaluation of voltmeter insertion error.

TABLE III Magnitude of Insertion Error Resulting from Voltmeter Resistance Being *n* Times the Source Resistance

n	Insertion error (%)
10	9.1
100	0.99
1,000	0.1
10,000	0.01
100,000	0.001

0.1%, which is 10 times greater than the meter's specified tolerance on the reading, and so an erroneous value for the voltage at the test point could be obtained.

An analysis of the insertion effects of an ammeter in comparison with the resistance of a circuit will result in a similar table, except that to minimize errors, the ammeter resistance should be a small percentage of the circuit resistance.

6. Measurement Error

The total or limit of error in a measurement is the sum of the errors from all the above and all must be considered before any can be considered insignificant in a particular measurement.

II. ELECTRICAL PARAMETER MEASUREMENT

Although in electrical and electronic engineering (and many other disciplines), it is necessary to determine the values of electrical parameters per se, it is frequently necessary to use electrical parameters to establish the value of a nonelectrical quantity such as temperature, position, or speed. This results in the need to measure a very wide range of electrical signals and quantities, which are categorized in the following sections.

A. Voltages

In measuring the amplitude of a voltage it may be first necessary to scale and change its amplitude and form to one suitable for the measuring device, for example, to reduce its amplitude from 100 V to 100 mV and convert it from an alternating to a direct voltage.

The method used to scale or reduce the amplitude of direct and low-frequency alternating voltages is commonly a form of resistance divider. The divider in Fig. 13a is generally used with electromagnetic analog instruments and has an input resistance that varies from range to range; that in Fig. 13b is frequently used in digital multimeters (and some electronic analog meters) and provides a fixed value of input resistance. When only alternating voltages have to be scaled, reactive dividers can be used and these may be capacitive or inductive. The inductive form can, with careful construction, provide a very precise voltage division. For example, a decade ratio accuracy (uncertainty) of 1 part in 10^7 (0.00001%) over a frequency range from 100 Hz to 20 kHz is obtainable with only moderate expense and care. Consequently such dividers are of considerable use in calibration work.

The scaling of power line voltages for protection and power measurement is performed for voltages up to 66 kV

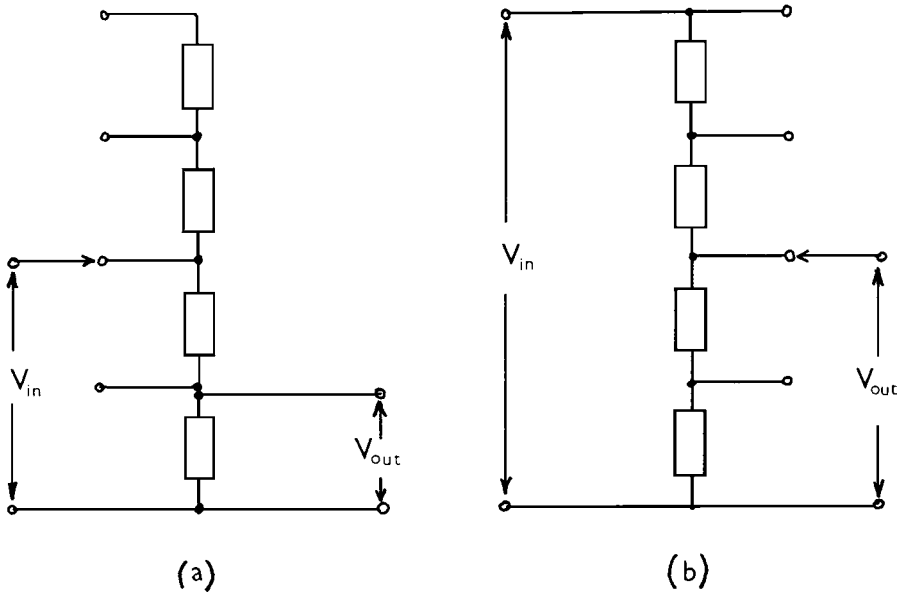


FIGURE 13 Resistive voltage dividers (a) for analog instruments, (b) for electronic/digital instruments.

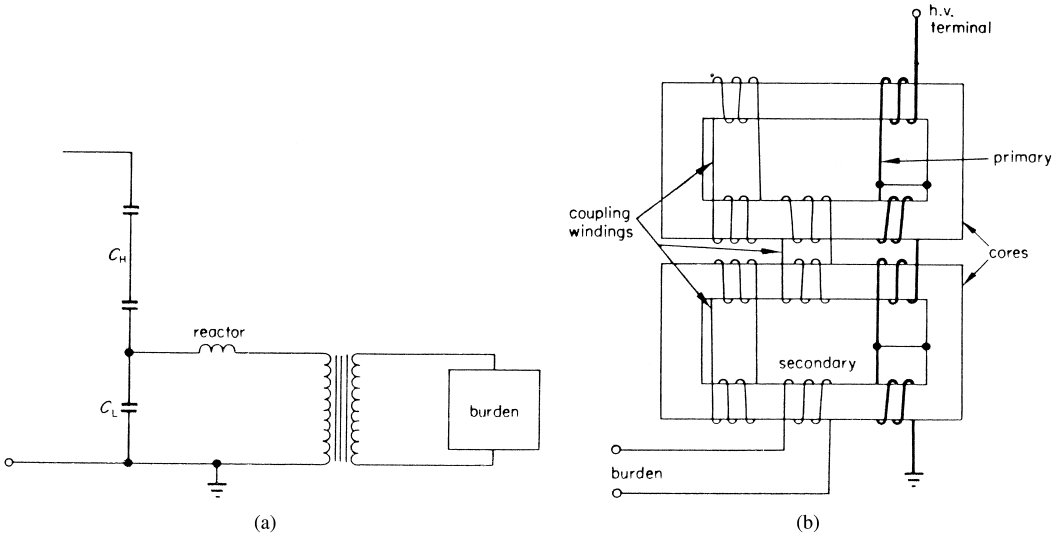


FIGURE 14 Voltage transformers (VT): (a) Capacitor VT and (b) cascade connected VT. [From Gregory, B. A. (1981). "An Introduction to Electrical Instrumentation and Measurement Systems," Macmillan Education, Hampshire, England.]

using voltage transformers that have a double-wound construction similar to that of a power transformer. For higher voltages, instrument transformers incorporate a capacitive divider (Fig. 14a) or use a cascade construction (Fig. 14b). All these voltage transformers will have phase and ratio errors due to the imperfections of the core material and the winding losses. Figure 15 shows a simplified phasor diagram for a voltage transformer, from which it is apparent that the phase angle error is the small angle between the primary and secondary phasors, and the ratio error is

$$\left(\frac{k_n U_s - U_p}{U_p}\right) \times 100\%.$$

The load on instrument transformers is of consequence as this will affect their errors in measurement. Consequently, the rating of an instrument transformer is given in terms of a specific load or *burden* and is normally quoted as a number of VA (voltamperes). For example, a voltage transformer rated 10 VA with a secondary voltage of 110 V should be operated with an instrument circuit that draws 10/110 A (or 0.09091 A). Thus, when the secondary voltage is 110 V, the total resistance of the load across the secondary winding should have a value of 1210 Ω.

Knowledge of these operating conditions is of consequence because the magnitude of the errors affects the measurement of power supplied (and charged) to a consumer.

The input circuitry of an instrument used to measure alternating voltages is frequently a combination of an amplifier and an R–C divider. The amplifier is necessary to increase the sensitivity of the measuring circuits or device and the divider to reduce large incoming signals to an appropriate magnitude. When a combined resistance and capacitance (R–C) divider is used (Fig. 16) providing the resistances and capacitors are of the correct proportions; the division of voltage is independent of frequency. Mathematically,

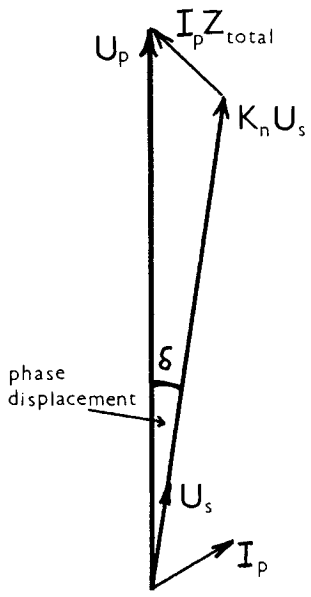


FIGURE 15 Phasor diagram of voltage transformer.

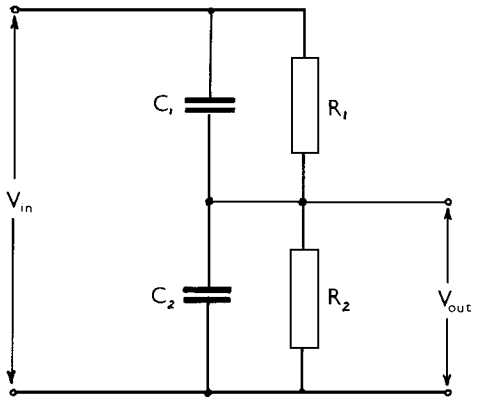


FIGURE 16 Resistance–capacitance voltage divider.

$$\frac{V_{\text{out}}}{V_{\text{in}}} = \frac{\dfrac{R_2/j\omega C_2}{R_2 + 1/j\omega C_2}}{\dfrac{R_2/j\omega C_2}{R_2 + 1/j\omega C_2} + \dfrac{R_1/j\omega C_1}{R_1 + 1/j\omega C_1}}$$

$$= \frac{R_2/(1 + j\omega C_2 R_2)}{R_2/(1 + j\omega C_2 R_2) + R_1/(1 + j\omega C_1 R_1)}.$$

Now, if R_1C_1 and R_2C_2, the time constants of the two halves of the divider, are made equal to T (by adjusting C_2, say), then

$$\frac{V_{\text{out}}}{V_{\text{in}}} = \frac{R_2/(1 + \omega T)}{R_2/(1 + \omega T) + R_1/(1 + \omega T)} = \frac{R_2}{R_1 + R_2}.$$

This shows that the voltage division is (theoretically) the same for all frequencies from zero (dc) to infinity. In practice, the imperfections of components result in an upper limit to the operating frequency of between 50 and 100 MHz.

While some forms of electromechanical instruments, notably those using an electrodynamometer movement, have a deflecting torque proportional to current squared and can thus provide a reading for zero frequency (dc) and alternating (20–200 Hz) signals, the vast majority of voltage-measuring analog and digital instruments can only be used to measure a zero frequency or direct voltage. Thus, for these voltmeters to be used for measuring alternating signals it is necessary to incorporate an ac-to dc-converter into the instrument.

The form of converter commonly used in analog multimeters has for many years been the full-wave or bridge rectifier. This provides a deflecting current that is dependent on the mean of the rectified value (Fig. 17) and results in a "mean-sensing" instrument. Since the rms value (dc or heating equivalent) is the required value, nearly all mean-sensing instruments are scaled to indicate the rms value, on the assumption that the signal being measured is a pure sine wave. Note that the quotient given by the rms value divided by the mean value is termed the *form factor* and equals 1.111 for a pure sine wave. A limitation on the use of the simple bridge arrangement is that a minimum voltage of 400 mV is required before the rectifying diodes conduct, which imposes a sensitivity limit that is often unacceptable in contemporary usage. The problem is overcome in electronic analog and digital instruments by incorporating diodes in an amplifier circuit as indicated in Fig. 18a, but the resulting arrangement is still mean-sensing.

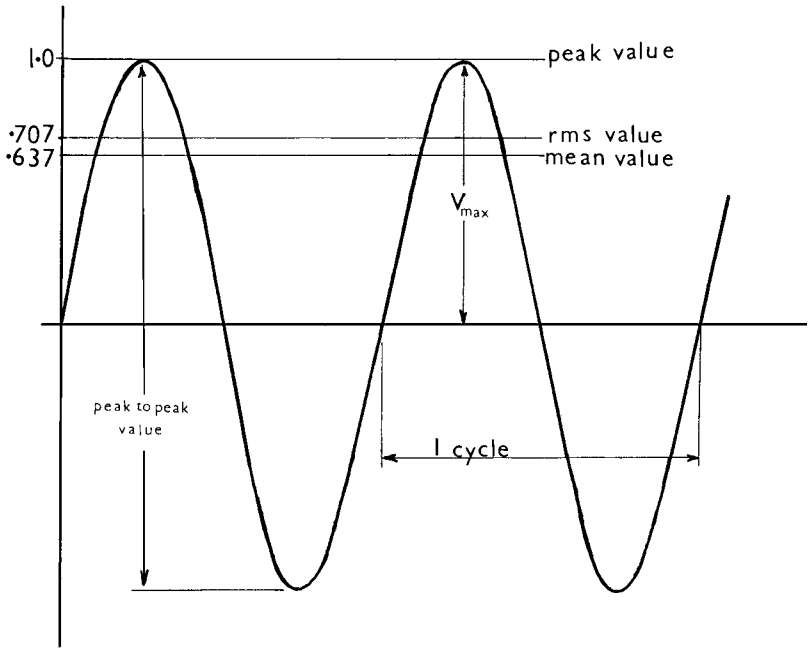


FIGURE 17 Mean, peak, and rms values of a single-frequency sine wave.

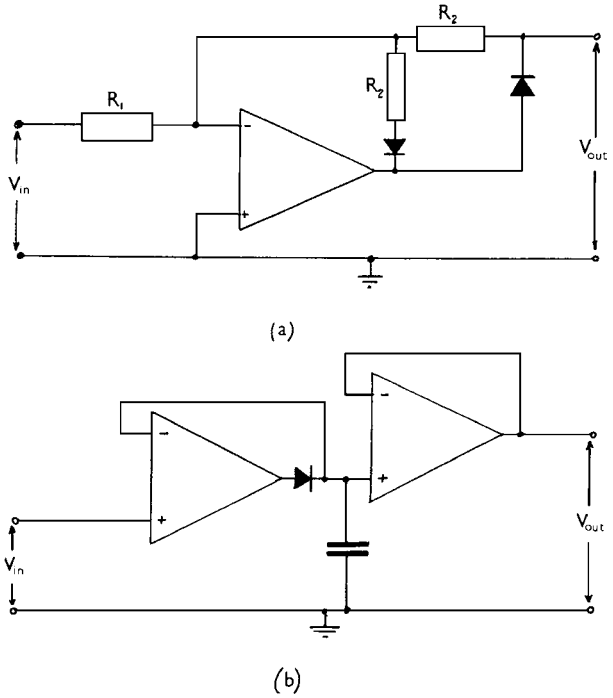


FIGURE 18 Ac-to-dc conversion: (a) Mean sensing and (b) peak sensing.

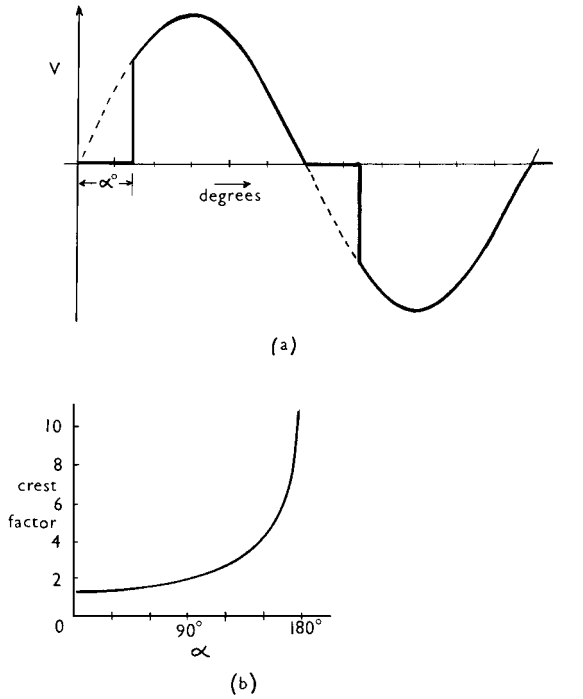


FIGURE 19 Sine wave switched off for α^{o}; (b) relationship between crest factor and α.

Should the peak value of voltage wave be required, a circuit of the form in Fig. 18b can be used to detect (and hold) the peak value. To provide an rms-sensing instrument, more complex electronic circuitry must be used, such as multiplier circuits, and to gauge the limitations of the electronic converter in coping with distorted waveforms, the *crest factor* (ratio of peak to rms value) should be specified for a "true" rms-sensing instrument. Since a number of voltage control applications now use a switched sine-wave technique in which the sine wave is switched off for part of the cycle (Fig. 19a) it is of interest to see the effect of varying the "off" portion (α^{o}) on the crest factor (Fig. 19b). Specified crest factors for digital multimeters vary from 3 to 10 and should be used accordingly, i.e., when there is a large amount of distortion in the signal, a meter with a large crest factor must be used. Each of the components in the input circuits that are subjected to alternating voltages will be affected by frequency. The result of this is that a specification will contain lower and upper frequency limits, typically 20 Hz at the low end and 5–30 kHz at the high end. This makes consideration of the specification essential, as operating outside of these limits will result in a (large) unknown error being added to the measurement.

Many mains-powered instruments have one input terminal that is connected directly (or via a small resistor) to the mains-supply ground. This grounded input terminal imposes a severe restriction on the use of an instrument, particularly when small signals are to be measured. Any voltage that exists between the ground (or chassis) of the instrument and the signal source will be added to the measurand, a point that is worth remembering and checking when using an oscilloscope, by selecting the maximum sensitivity and connecting the oscilloscope input to the source ground, thereby measuring any signal present at the source ground.

The input-ground problem is overcome in most contemporary electronic instruments by having an input that is "floating." This means there is no direct connection between the input terminals and ground, for the measurement circuits are insulated (i.e., isolated) from the ground connection. This situation is represented in Fig. 20, in which

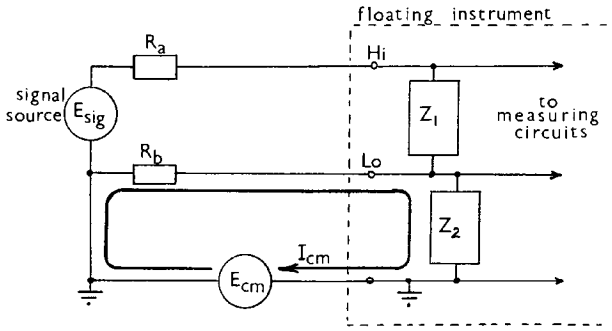


FIGURE 20 Simplified equivalent circuit of instrument with floating input.

Z_2 represents the impedance of the insulation and will have a large value (typically 10^8 Ω).

To measure small signals remote from the instrument it is desirable to use an instrument with a "guarded" input (Fig. 21), for this introduces additional insulation between the measuring circuits and ground and, more importantly, provides a path that diverts error-producing currents away from the signal-sensing circuits.

So that the insulation/isolation of the measuring circuits to ground of one meter can be compared with those of another meter, a convention has been established. This concerns a hypothetical situation represented by Fig. 22, which provides for the evaluation of the common mode rejection ratio (CMRR), which equals $20 \log_{10}(E_{cm}/E_e)$ dB, where E_{cm} is the common mode voltage and E_e the error introduced in the reading by E_{cm} when the unbalance resistor is 1 kΩ. Because there is a capacitive component in the insulation impedance, the ac CMRR should be specified at a particular frequency and will be less than the dc value.

Note: As it is unlikely in practice that the unbalance resistor will be 1 kΩ, the expression for the CMRR should be modified to

$$20 \log_{10}\left(\frac{E_{cm}}{E_e \times (1\ \text{k}\Omega / R_{unbal})}\right) \text{dB}.$$

When measuring small direct voltages, a frequent problem is the effect of alternating interference superimposed on the measurand. The ability of an instrument (in its direct-voltage measuring mode) to reject this interference is its normal (or series) mode-rejection ratio capability. Numerically,

$$\text{NMRR} = 20 \log_{10}\left(\frac{E_{nm}\text{peak}}{E_e\text{peak}}\right) \text{dB}$$

at a specified frequency,

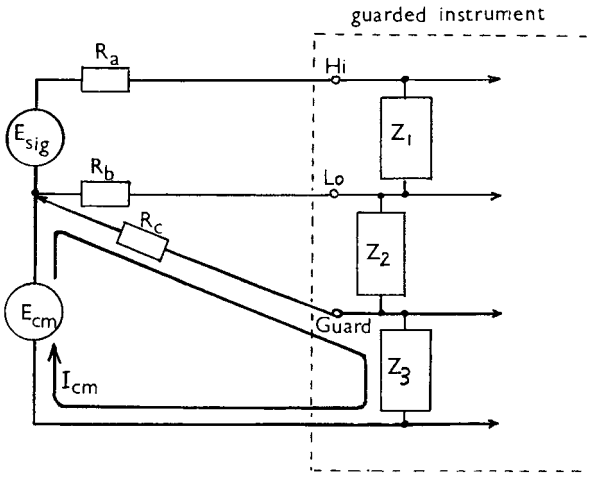


FIGURE 21 Simplified equivalent circuit of instrument with guarded input.

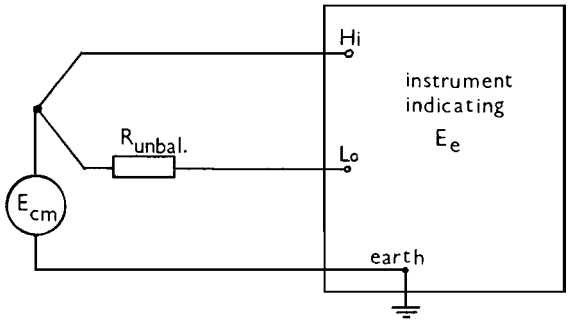


FIGURE 22 Common-mode rejection ratio evaluation circuit.

where E_{nm} is the normal mode voltage and E_e the resulting reading error, which is usually evident as a fluctuation in the displayed value.

B. Currents

While the majority of analog instruments are current-measuring devices, that is, the pointer deflection is dependent on the flow of current through the instrument, the current-carrying capacities of the instrument's coils are generally limited to milliamperes, or at the most a few amperes. The exception to this is some specially constructed electrodynamometer and moving-iron meters used for measuring currents of the order of 100 A. Since currents from picoamps to kiloamps are in use, current scaling is as important and useful as voltage scaling.

To facilitate the measurement of currents larger than the capacity of the conductors in an instrument, current shunts may be used. These provide a low-resistance path parallel to the measuring instrument and are in general a four-terminal arrangement (Fig. 23). In use, the measurand is connected to the current terminals and the measuring instrument to the potential terminals.

The applications of current shunts lie in the extension of ammeter ranges (where the voltage drop across the shunt drives a proportion of the current through the instrument) for the measurement of low-voltage power, the recording of current waveforms, and the measurement of current by voltage-sensing instruments, it being appreciated that with the increasing use of electronic measuring instruments (which invariably have high input resistances), the measurement of current as a voltage drop across a known resistance is becoming a much used method. While the type of shunt illustrated in Fig. 23 is intended for measuring medium to large direct currents, the use of shunts is not limited to dc; if a shunt is nonreactive, it can be used for alternating current measurements. The principal disadvantages of the shunt are its power consumption and the fact that the metering connections must be operated at

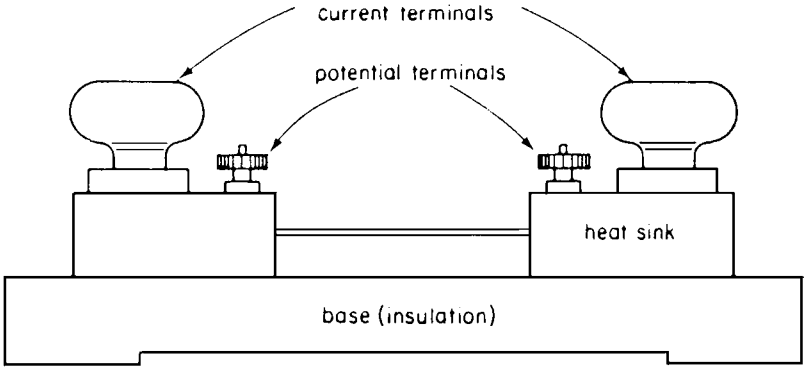


FIGURE 23 Four-terminal direct current shunt. [From Gregory, B. A. (1981). "An Introduction to Electrical Instrumentation and Measurement Systems," Macmillan Education, Hampshire, England.]

the same voltage to ground as the current-carrying line; the latter, for safety reasons, limits shunt use to low-voltage applications.

The voltage drop method of measuring current, used for the majority of digital multimeter current ranges, causes a voltage drop of between 0.2 and 2 V in the measured circuit. In many digital multimeter specifications this voltage drop is termed the *burden*, terminology originating from instrument transformer specifications (see Section II.A). In situations where this burden causes an unacceptable insertion error, such as in the measurement of very low currents, feedback ammeters should be used because their voltage drop is 1 mV or less. This arises because the voltage drop V induced in the measured circuit is equal to the voltage drop across the meter V_o (Fig. 24) divided by the amplifier gain. Since the amplifier gain is large, the resulting V_m becomes very small.

A current-divider arrangement that can be incorporated into a galvanometer as a sensitivity control, and is thus useful when such an instrument is used as a null detector, is the universal shunt. A much wider application of this device is to provide the various current ranges of an analog multimeter. Figure 25 shows the basic arrangement for a universal shunt of resistance R shunting a meter of resistance R_m. With the switch set to position a the meter current $I_m = IR_1/(R + R_m)$, and if $R \gg R_m$, $I_m = I/1000$. Similarly, at position b, $I_m = I/100$; at position c, $I_m = I/10$; and at d, $I_m = I$. By suitable selection of resistance values any division of current can be obtained, and even if R is not much larger than R_m the ratio of division is retained, although the damping of the meter movement will be affected.

The current transformer (CT) overcomes the power loss and circuit isolation problems of the current shunt, but like the VT, it introduces ratio and phase-displacement errors. The construction of a current transformer (Fig. 26) is different from that of a power transformer, although the basic theory of all transformers is the same, that is, (a) the voltage induced in a transformer winding is proportional to $N\Phi_m f$ (where N is the number of turns in the winding, Φ_m is the flux in the core, and f is the frequency of operation) and (b) the ampere-turns of the windings balance.

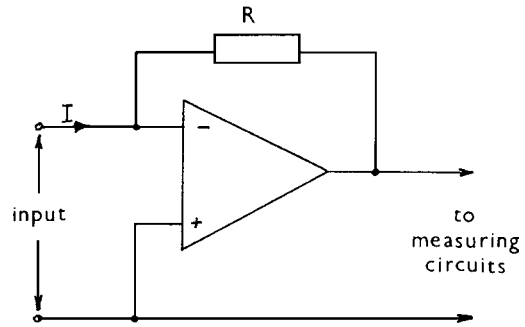


FIGURE 24 Low-impedance electronic ammeter circuit.

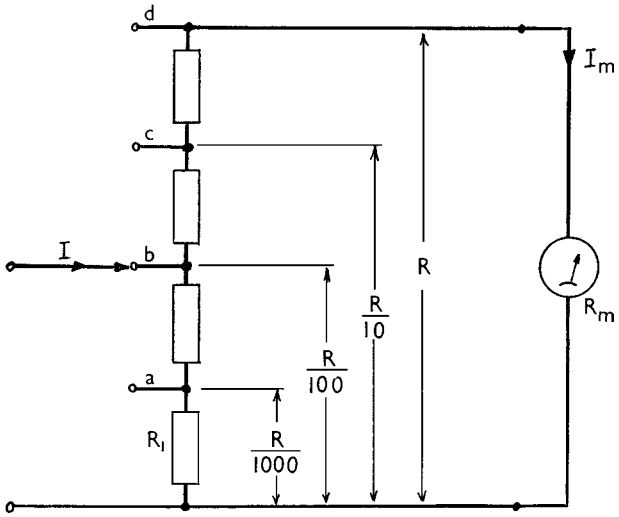


FIGURE 25 Circuit arrangement of a universal shunt.

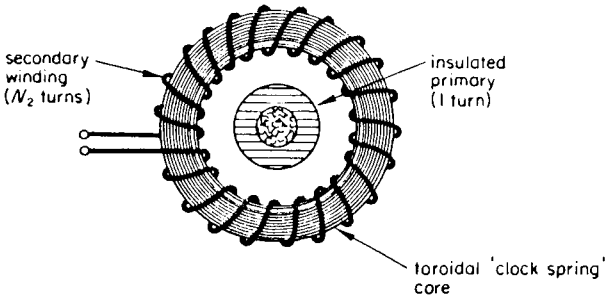


FIGURE 26 Current transformer with a bar primary. [From Gregory, B. A. (1981). "An Introduction to Electrical Instrumentation and Measurement Systems," Macmillan Education, Hampshire, England.]

It is the latter feature that is exploited in the design of a current transformer, in which the no-load or magnetization current is made as small as possible to ensure that the current ratio approaches the inverse of the turns ratio. The magnetization current is minimized by using a toroidal core wound from a continuous strip of high-permeability steel. The primary winding is usually a single turn (referred to as a bar primary) but may consist of a tapped winding to produce a multiratio CT. The secondary winding, rated at either 1 or 5 A, will have many turns tightly wound onto the core.

To investigate the current error (ratio error) and phase displacement for a current transformer, consider the simplified phasor diagram in Fig. 27. It can be seen that the difference in magnitude between the primary current I_p and the secondary current multiplied by the rated transformation ratio ($I_s K_n$) is dependent on the amount of the primary current used to energize the core, which must therefore be kept to a minimum. It should be noted that a CT is operated with a load consisting of an ammeter that almost short-circuits the secondary terminals; should this

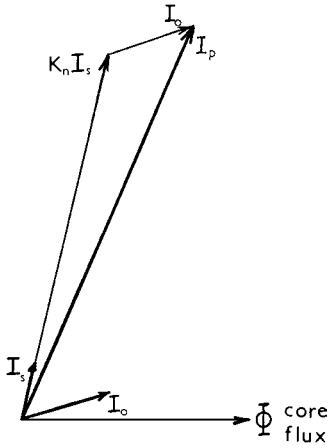


FIGURE 27 Simplified phasor diagram of a current transformer (not to scale).

load be open-circuited, the whole of the current in the primary winding will become the energizing current, causing magnetic saturation of the core (which is detrimental to its magnetic properties) and causing a large, peaky voltage to occur at the secondary terminals (which may result in failure of the interturn insulation, apart from being dangerous for the operator). Thus, when connected in a circuit, a CT should *never* be open-circuited.

The secondary load on a current transformer is termed the *burden* and is expressed in ohms or volt-amperes at the rated current. A commonly used value is 15 V-A, which for a CT with a secondary rated at 5 A requires a secondary terminal voltage of 3 V and an external (or load) impedance of 0.6 Ω for operation at the correct burden.

C. Power Measurement

The measurement of the power consumed by a load at any instant requires the determination of the product of the instantaneous values of the appropriate voltage and current, that is, $P = vi$. For the direct voltage and current, the power consumed is simply the product of the voltage across the load and the current through it, since the instantaneous and mean values are the same. However, this simple condition is not true for the load subjected to an alternating supply because of the continuously varying amplitude of a sine wave. Additionally, there will almost certainly be a phase difference between the applied voltage and the current flowing, as in Fig. 28.

The mean power over one cycle is

$$P = \frac{V_m I_m}{4\pi} \int_0^{2\pi} \cos\phi \, d\theta$$

from $v = V_m \sin\theta$ and $i = I_m \sin(\theta + \phi)$. Thus $P = (V_m I_m / \sqrt{2}\sqrt{2}) \cos\phi = VI \cos\phi$, where V and I are the rms (dc equivalent) values of voltage and current, respectively.

For many years the most common method of measuring ac power has been to use an electrodynamometer instrument. This analog meter has a deflecting torque proportional to the product of the currents flowing in the fixed and moving coil systems. Figure 29 is a diagram of the electrodynamometer instrument, which has always been a relatively expensive instrument to construct. When measuring power, the moving coil is connected via a series resistor across the load, so the current through it is proportional to the load voltage and the fixed (current) coils are connected in series (or via a current transformer) with the load. Due to the power consumption of the electrodynamometer wattmeter, consideration of the error resulting from the connection method must be made. If the voltage coil is connected on the supply side of the current coil, the deflecting force proportional to voltage is larger

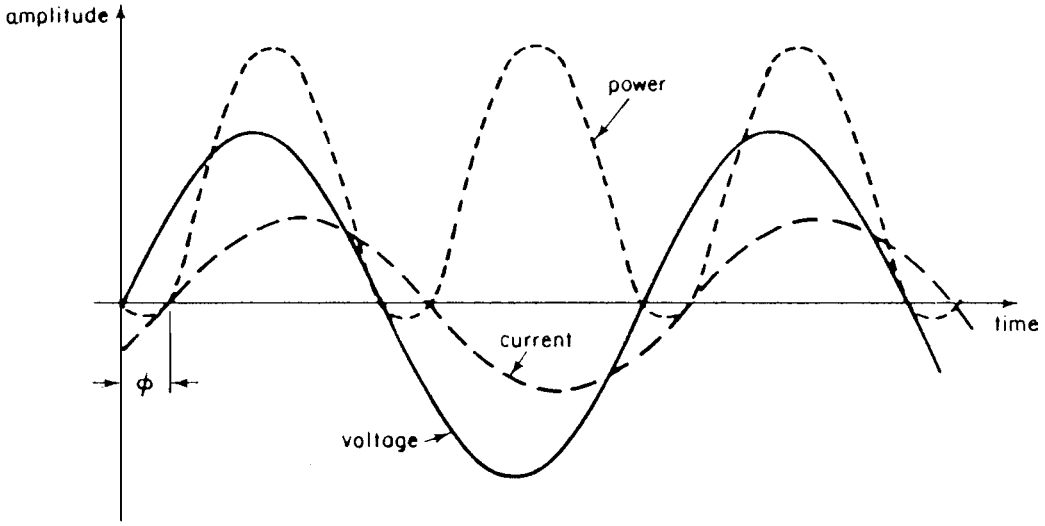


FIGURE 28 Voltage, current, and power waveforms in a reactive circuit.

than it should be, whereas when the voltage coil is connected on the load side of the current coil, the deflecting force proportional to current is erroneously large. In the former case, $I_L^2 R_C$ should be subtracted from the meter reading, whereas for the latter case, V_L^2/R_V must be deducted, where I_L and V_L denote the load current and voltage, respectively, and R_C and R_V are the resistances of the wattmeter current and voltage circuits, respectively.

The electronic wattmeter is beginning to supplant the electrodynamometer instrument since it overcomes the loading disadvantage of the electrodynamic instrument. However, particularly when the display is a digital one, it is a moderately expensive instrument. The technique used to sense the power is essentially the same, deriving signals proportional to the (instantaneous) load current and voltage. The determination of the power is either by an analog process utilizing an electronic multiplier arrangement or by a digital process that makes many simultaneous samples of both voltage and current during one cycle of the system frequency and then computes the power from the mean of the instantaneous power values. In either case, because electronic sensing is used, the input impedance of the voltage circuit will be greater and that of the current circuit lower than their counterparts in the electromechanical instrument. Example input impedance figures for the two types of instrument are 5 kΩ/V and <60 mΩ for an analog electronic wattmeter, compared with 45 Ω/V and 1 Ω for the electromagnetic instrument, where the respective values are for the voltage and current circuits of the two types of instrument. These improved input impedance values result in an instrument that is simpler to use; its connection between a power supply and a load is shown in Fig. 30. Additionally, it is usually unnecessary to calculate the correction due to the loading of the meter when measuring the power taken by a load. Further advantages of the electronic wattmeter are that the range of power that can be measured and the operational bandwidth are usually far greater than those of an electromechanical instrument.

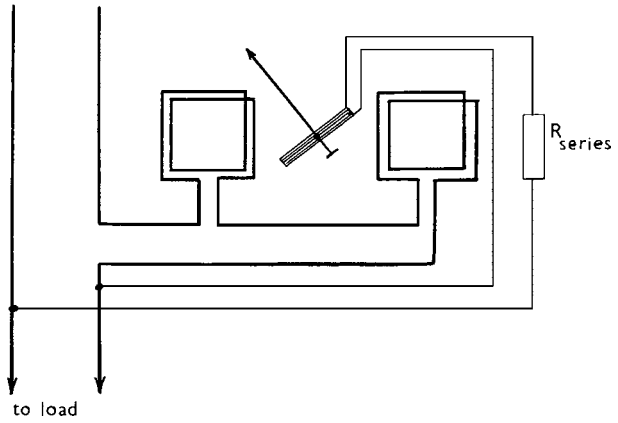


FIGURE 29 Diagrammatic representation of the connections in an electrodynamometer wattmeter.

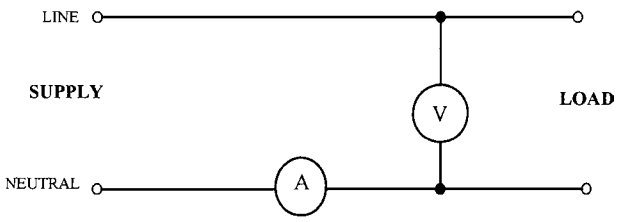


FIGURE 30 Connection circuit for an electronic analogue wattmeter.

D. Resistance Measurement

Probably the circuit most widely used for many years for the measurement of resistance is that attributed to Sir Charles Wheatstone. Whereas he was an accomplished scientist, making many contributions to science and engineering, he only evaluated the application of the bridge circuit that bears his name. The circuit, shown in its basic form in Fig. 5, was originally devised by S. H. Christy in 1833. This circuit has been used in various forms in a proliferation of applications from the Kelvin double bridge used for low-resistance measurement (Fig. 31) to the unbalanced arrangement used in many transducers.

The balance equation for the basic circuit of Fig. 5 is X (the unknown) $= SP/Q$, in which S is a known variable arm (decades of resistance values); the arms P and Q are often referred to as the ratio arms. If switchable values of 10, 100, 1000, and so on are provided in these arms, a wide range of unknown resistors can be measured.

For the unbalanced bridge shown in Fig. 32, the output voltage when one arm is varied by a small amount $\pm\delta R$ is $V_{\text{out}} = V_{\text{in}} \cdot \delta R/(4R \pm 2\delta R)$; see Section IV.B. The balance equation of the Kelvin double bridge has been given in Section I.F.

The method traditionally used for resistance measurement in multimeters and ohmmeters has been an application of Ohm's law. In operation, a known current from a constant-current source is passed through the unknown resistor and the voltage drop across the resistor measured so the value of the unknown resistance may be calculated. However, by suitable scaling the voltage drop can be displayed as a resistance value. For example, passing 1 mA through a resistance of 1.523 kΩ would produce a voltage drop of 1.523 V across it. Measuring this voltage drop with a high-impedance (digital) voltmeter and using the circuit arrangement in Fig. 33a provides a two-terminal

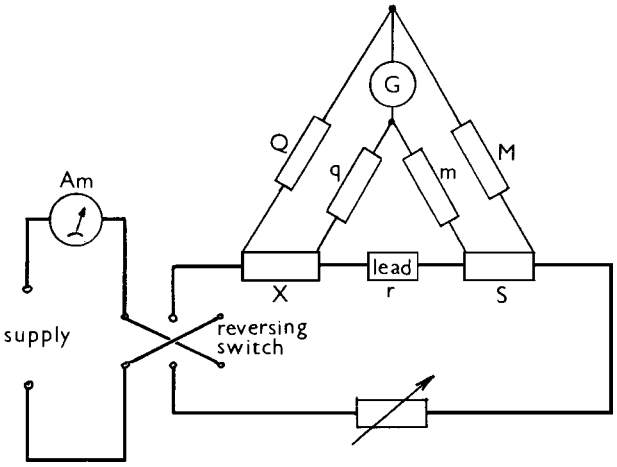


FIGURE 31 Kelvin double-bridge circuit.

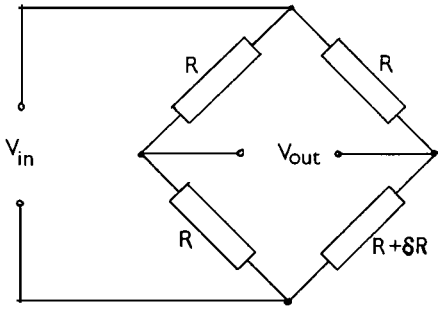


FIGURE 32 Unbalanced Wheatstone bridge.

measurement which is suitable if the display resolution is greater than the lead resistance r (say, 0.1 Ω), while Fig. 33b is appropriate when this is not the case (i.e., for low-resistance measurement).

With the inclusion of calculation capability in digital multimeters, many manufacturers have moved to a ratio process for the resistance measurement function; this technique is outlined in Figs. 34a and 34b for the two- and four-terminal measurement situations, respectively. The main advantages of this arrangement are the measurement error reduction due to the use of a ratio process and the replacement of a precision current source by a less expensive voltage source, which, while it needs to have good stability, does not require a known constant amplitude. The unknown resistance is evaluated from the expression $R_x = R_{\text{ref}} V_x / V_{\text{ref}}$, where V_x and V_{ref} are the voltages across the unknown and R_{ref} respectively. The value of R_{ref} (a known stable reference resistance) is held permanently in a memory and V_x and V_{ref} are measured

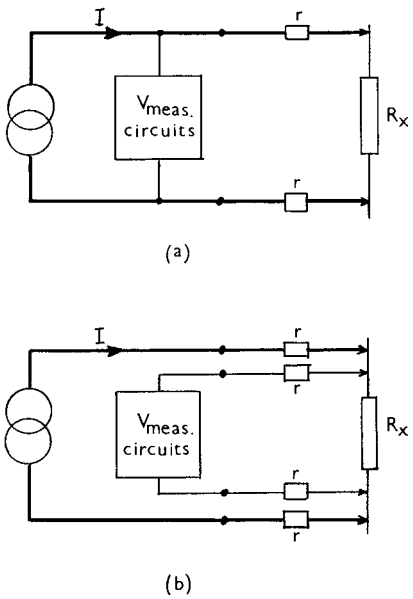


FIGURE 33 Resistance measurement (Ohm's law) arrangements: (a) Two terminal and (b) four terminal.

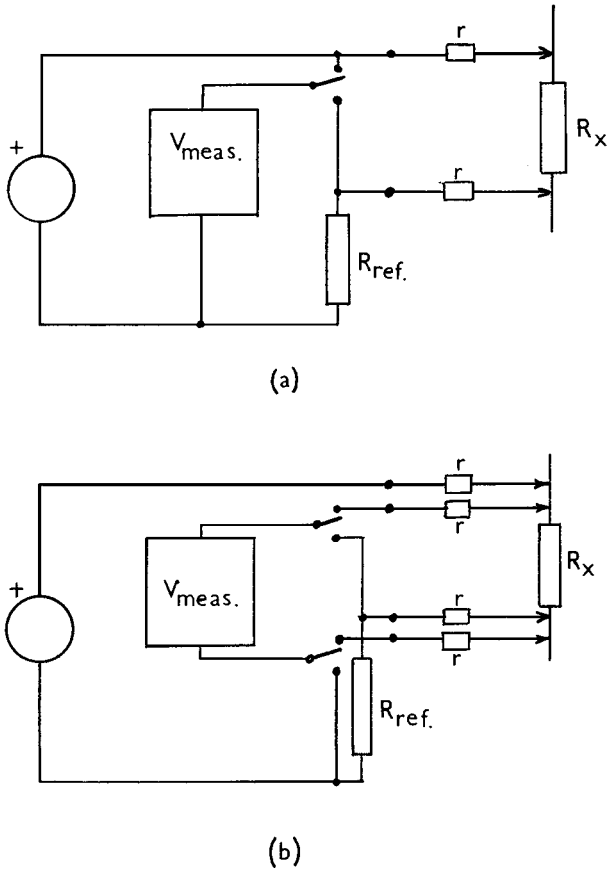


FIGURE 34 Resistance measurement (ratio technique): (a) Two terminal and (b) four terminal.

sequentially prior to the calculation and display of the value of the resistance being measured. Some multimeters can measure and store an initial value (e.g., the lead resistance) and then display measurands less the initial value, a process that eliminates the need to make four-terminal measurements.

The expansion of automated testing and the need to measure the components of assembled circuits at stages in their manufacture have produced a requirement for in-circuit measurement. All networks of circuit elements can be reduced to a star or delta arrangement (sometimes referred to as T and π, respectively). In the star configuration there is no problem since one end of the component is disconnected and no alternative circuit path exists. However, in the delta situation a permanent parallel path exists around the unknown, and while it might appear possible to open-circuit the parallel path, in practice such a step would not be practicable due to time considerations and the possibility of causing damage to the components or the board on which they are mounted. The measuring techniques that have evolved to solve this problem can be divided into two variations of the same idea in which a

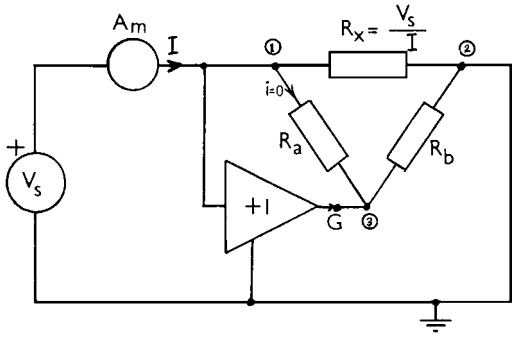


FIGURE 35 In-circuit component measurement, guard raised to node potential.

"guard" point in the parallel circuit is forced to the same potential as one end of the unknown resistor. Figure 35 shows the guard point (3) raised to the potential of node 1 so that no current flows through R_a, and $R_x = V_S/I$. Figure 36 shows the guard point and node 2 effectively at ground potential; in consequence, zero current flows through R_b. In both cases, the technique will only operate over a range of R_x values dependent on the magnitudes of R_a and R_b. Generally, the limitations are expressed as (i) the guard ratio $R_x/[R_aR_b/(R_a + R_b)]$, which can have values between 100 and 1000, and (ii) the minimum magnitude of R_a and R_b (typically 200–500 Ω).

E. Capacitance and Inductance

These components have relevance only in alternating current circuits. At dc a capacitance appears to be an open circuit (apart from a small leakage current through the resistance of the dielectric), and an inductor appears to be a small resistance (that of the coil conductor). However, on connection or removal of the dc supply, the effects of capacitance and inductance are apparent due to the change in the energy stored in the device.

In ac circuits the properties of both inductors and capacitors are affected by frequency, and while these effects

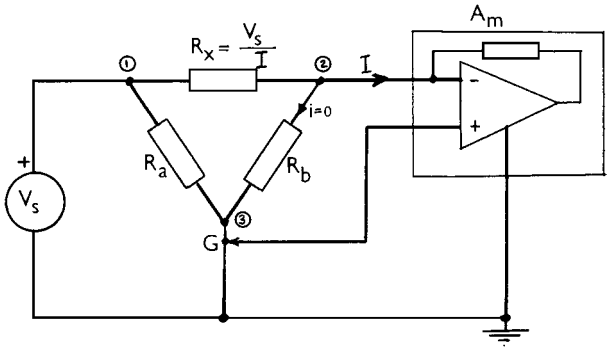


FIGURE 36 In-circuit component measurement, virtual ground guard.

are not generally apparent in capacitors until frequencies above a few megahertz are reached, the properties of indictors may begin to show changes at a few kilohertz. In making measurements on inductors and capacitors, it is therefore relevant to record the frequency at which the measurement has been made and, if possible, to make the measurement at the intended frequency of operation. As with resistance measurement, bridge methods have been widely used for many years for both inductance and capacitance measurement.

In the operation of a bridge circuit, both resistive and reactive components of the unknown must be considered. For while small capacitors made using modern dielectric materials are effectively pure reactances at low and medium frequencies, as the (parallel) leakage resistance has a very high value, this was not the case 20 or so years ago. Nor is it true for inductors or for large-value capacitors. The need to know these component parts led to a proliferation of bridge circuits based on the Wheatstone circuit. Well-used examples of ac bridges are those attributed to Maxwell, Owen, and Heaviside for inductance measurement and to Wein, De Sauty, and Schering for capacitance measurement. The last of these (Fig. 37) is still in use because of its high-voltage applications.

An alternative to bridge methods and suitable for use at frequencies from 1 kHz to 300 MHz was the Q-meter, a device which is no longer appears in the catalogues of instrument manufacturers but is likely to still be in use in some locations, as it capable of measuring small capacitors over a wide range of frequencies. The Q of a component indicates its quality and is defined as 2π times the ratio of energy stored to energy lost per cycle. Numerically, this is the ratio of reactance to resistance at the frequency under consideration. The reciprocal of Q (that is, $1/Q$) is the dissipation factor D, both terms being widely used for the purpose of indicating the quality of inductors and capacitors.

The Q-meter operates on the principle of creating a resonant condition between an inductor and a capacitor. Figure 38 shows a simplified Q-meter circuit in which

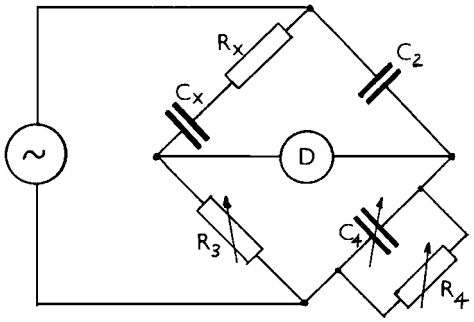


FIGURE 37 Schering bridge circuit.

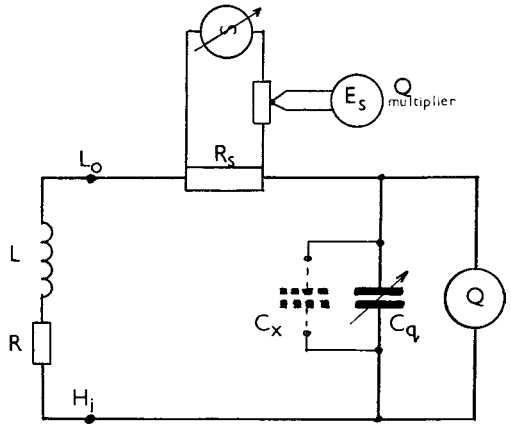


FIGURE 38 Q-meter, simplified circuit.

resonance can be created by variations of the source frequency or by adjustment of C_q. In either case, resonance is indicated by a maximum of the voltage E_q across C_q when

$$\frac{E_q}{E_s} = \frac{1}{\omega_0 CR} = \frac{\omega_0 L}{R} = Q \qquad \text{and} \qquad Q = \frac{1}{D}.$$

The operator skill required to perform reliable measurements using bridge circuits and the Q-meter was appreciable and the contemporary desire for digital display, the availability of cheap computing power, and improvements in electronic circuit reliability have led to the replacement of these instruments by impedance or *LCR* meters. In operation, either an Ohm's law or a ratio process is used, the latter being an inherently more precise technique, but requiring a more sophisticated switching process.

The basic arrangement of an *LCR* meter is shown in Fig. 39, in which a current at a known frequency is supplied

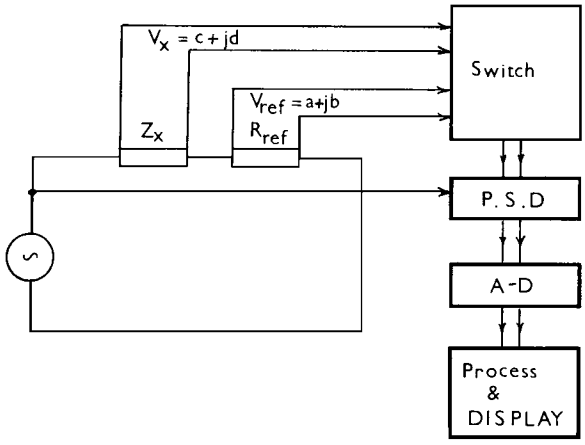


FIGURE 39 Basic *LCR* meter circuit.

to the unknown or "device under test" (DUT) while connected in series with a known or reference resistance. In the ratio form of operation, the voltage across the reference (V_{ref}) and then that across the device under test (V_x) are fed in turn to the phase-sensitive detector (PSD), so that the in-phase and quadrature components of both voltages with respect to a reference signal can be determined and fed to an analog-to-digital (A–D) converter. The processor section then contains the components of V_{ref} as $(a + jb)$ and those of V_x as $(c + jd)$, from which the required parameters of the unknown can be evaluated and displayed. For example, the series components of an inductor can be displayed from $Z_x = V_x \cdot R_{\text{ref}}/V_{\text{ref}} = R_x + jX\omega L_x$, or

$$R_x = R_{\text{ref}}\frac{ac + bd}{a^2 + b^2} \qquad \text{and} \qquad \omega L_x = \frac{ad - bc}{a^2 + b^2}.$$

Should the parallel components of a capacitor be required, then these can be obtained from $Y_x = V_{\text{ref}}/R_{\text{ref}}V_x = G_x - j\omega C_x$, or

$$R_{xp} = R_{\text{ref}}\frac{c^2 + d^2}{ac + bd} \qquad \text{and} \qquad \omega C_x = \frac{ad - bc}{(c^2 + d^2)R_{\text{ref}}}.$$

In addition to these basic quantities, *LCR* meters generally are capable of displaying impedance Z and its phase angle, resistance and series reactance, admittance Y, conductance, parallel susceptance, Q, and D.

III. TIME-DEPENDENT QUANTITIES

The measurement of electrical signal variations with time can in some situations be of considerable significance, for example, the time taken from the initiation of a process to its completion. For alternating (sine wave) signals, frequency is usually the quantity of consequence, while for step changes in signal level the rise time and overshoot are generally of importance. Consequently, it is often desirable to monitor or display the variations with time that are occurring in a signal so that these variations are made visible. The type of variation and the rate at which a signal varies with time dictate the form of measuring instrument used to display and/or record the signal waveform. Furthermore, if frequency measurement on periodic signals or the determination of a time interval is required, digital counters can be used to provide a precise result.

A. Waveform Monitoring

As a result of the use of TV and computers, the cathode ray tube (CRT) or monitor has become an accepted method of displaying information. In general, the electron-beam deflection system used in a cathode ray oscilloscope is electrostatic and that in a TV or computer monitor is electromagnetic, although the use of liquid crystal displays is becoming more widespread and has resulted in the production of some 'flat' oscilloscopes (Fig. 49). The advantage of electromagnetic deflection is that a large display area is obtained with a relatively short tube length; due to the inherent electromagnetic properties of coils, however, the operating frequency range of the deflection system is very limited. The advantage of electrostatic deflection is an increased frequency range, but to obtain a large display a long tube is required. The principle of operation of both systems is the same.

Figure 40 shows the arrangement of a cathode ray oscilloscope (CRO) with an electrostatic deflection system. Inside the CRT is an electron gun that projects a fine stream of electrons between the deflecting plates onto a phosphor-coated screen, where a luminous spot is formed. The focusing of this spot is controlled by the magnitude of the direct voltage applied to biasing electrodes. The quantity of electrons projected and hence the maximum writing speed of the oscilloscope is dependent on the voltage difference between the various electrodes in the electron gun.

To move the spot horizontally across the tube face or screen at a constant speed, a sawtooth waveform (Fig. 41) is applied to the X deflection plates. The rising slope of the sawtooth waveform is adjustable in selectable values

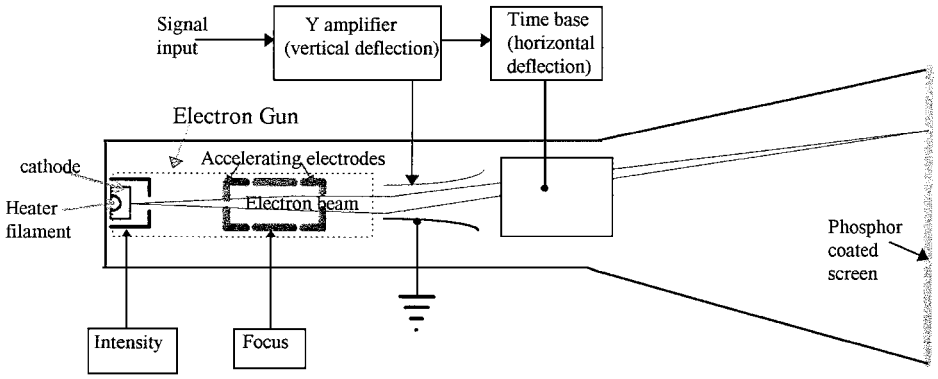


FIGURE 40 Schematic arrangement of cathode ray oscilloscope (CRO).

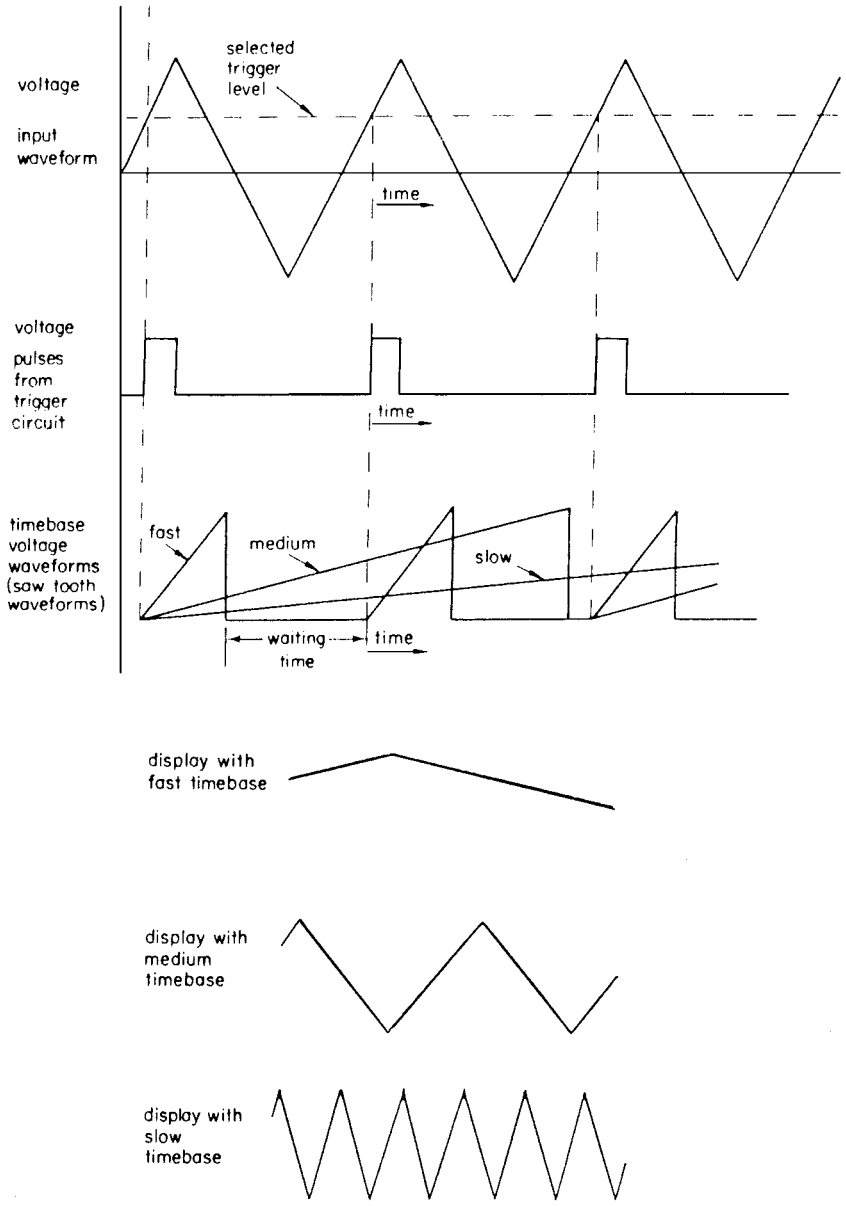


FIGURE 41 Effect of timebase selection on the waveform displayed on a CRO. [From Gregory, B. A. (1981). "An Introduction to Electrical Instrumentation and Measurement Systems," Macmillan Education, Hampshire, England.]

so that the spot can be deflected horizontally (from left to right) at known speeds, typically in a 1, 2, 5, 10 sequence from 1 sec/DIV (1 DIV = 1 division on the graticule or grid on the tube face) to 200 nsec/DIV, and is referred to as the time base. The falling edge of the sawtooth is made fast, so that the "flyback" (right to left deflection) speed is above the writing speed and hence not visible. To obtain a vertical displacement of the spot, the signal to be monitored is applied via an attenuator and amplifier combination to the Y plates. To obtain a stable display of a periodic wave, it is necessary to be able to trigger or start the operation of the time base at a selectable point on the measurand, as indicated in Fig. 41.

The Y or vertical deflection system usually has selectable values in a 1, 2, 5, 10 sequence to give deflections from 5 mV/DIV to 20 V/DIV for a general-purpose oscilloscope, or from 50 μV/DIV in some limited-bandwidth oscilloscopes for special applications. In the majority of oscilloscopes, one side of the vertical input socket is grounded since the input connector is commonly of the BNC type used with coaxial cable, the outer conductor or screen being "earthed." Between

the input socket and the vertical deflection system is a three-position switch (Fig. 42). In the dc position of this switch all components of the signal are conveyed to the deflection system; in the ac position any dc bias on an input signal is removed so that just the alternating component of a combined ac and dc signal can be observed. The GD (or GND) position provided on the input switch isolates the measurand from the input amplifier and connects the amplifier input but *not* the input terminal to ground. In this ground position the display on the tube face is a straight horizontal line that provides a zero reference line so that the presence of a direct voltage can be detected and measured.

In using an oscilloscope to perform measurements on a waveform, it should be remembered that the CRO display provides values that have, typically, ±3% tolerances. The greatest value of the CRO is the display of what is occurring to a signal at a point in a circuit, for example, if it has become distorted during its progress through the circuit. To determine the properties of a displayed wave, use is made of the graticule and the sensitivity settings on the *Y* amplifier and the time base. For example, from the display of a waveform such as that in Fig. 43a and with settings on the time base of 10 μsec/DIV and vertical amplifier of 20 mV/DIV, the amplitude and frequency can be determined as $6.8 \times 20 = 136$ mV peak-to-peak or 48 mV rms and the duration as 8 DIV $\times$ 10 μsec/DIV $= 80$ μ sec/cycle or 12.5 kHz since the peak-to-peak value of a waveform is $2\sqrt{2}$ times the rms value and the frequency of a wave is the reciprocal of its duration or period.

When monitoring square or pulse waveforms, the property usually of greatest importance is the rise time, defined as the time taken for the wave to change from 10% to 90% of its final value. To ease this type of measurement, the graticule has a pair of dotted lines across it at 10% and 90% of the distance between the top-but-one and the bottom-but-one horizontal lines of the graticule. By manipulation of the variable gain control of the vertical amplifier and vertical position control it is possible to obtain a display of the form shown in Fig. 43b; from this display and the time base setting, the rise time is rapidly established. For example, considering the wave in Fig. 43b, if the timebase is 1 μsec/DIV, the rise time is 2 DIV $\times$ 1 μsec/DIV, i.e., 2 μsec.

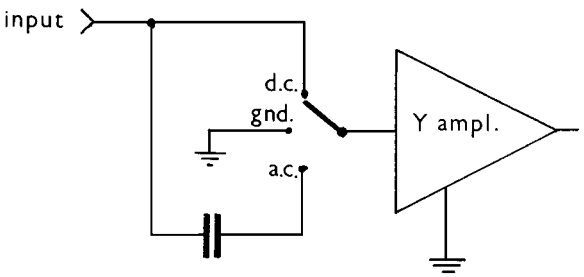


FIGURE 42 CRO input selection switch.

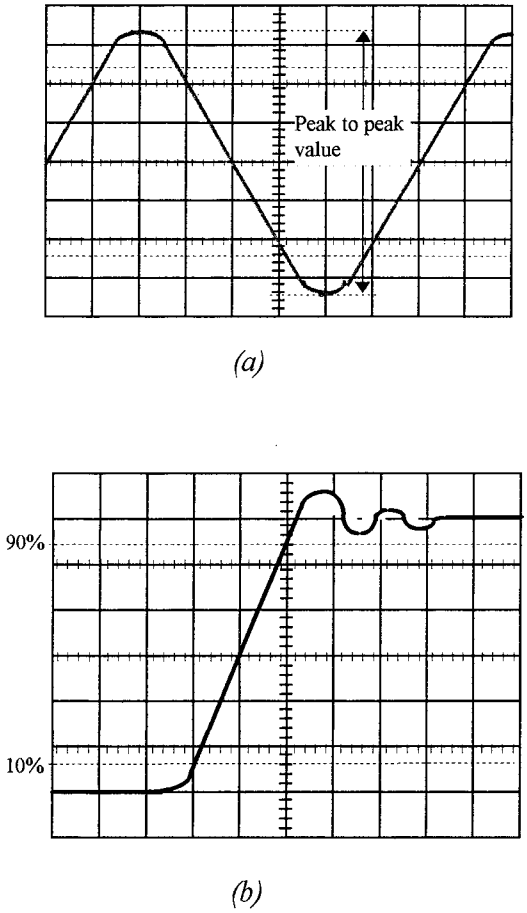


FIGURE 43 (a) Measurement of voltage from a CRO display; (b) rise time and overshoot measurements on the CRO display of a step function.

Of paramount importance when using an oscilloscope is an awareness of its bandwidth (or frequency range). This is specified as the -3 dB point, i.e., the frequency at which its amplitude response has fallen by 30% from the mid-frequency value. For example, if these values for a general-purpose oscilloscope are dc and 20 MHz, the frequency response curve is as in Fig. 44. Now, since the amplitude response at the -3 dB frequency is reduced by 30% (strictly, 29.3%), and this is likely to be an unacceptable error, it is necessary to consider at what frequency a response that is within, say, 5% of the mid-frequency would occur. Since the -3 dB frequency value is obtained from the expression

$$\frac{V_{\text{disp}}}{V_{\text{input}}} = \sqrt{\frac{1}{1 + \left(1/\omega_{3dB}^2 \times T^2\right)}},$$

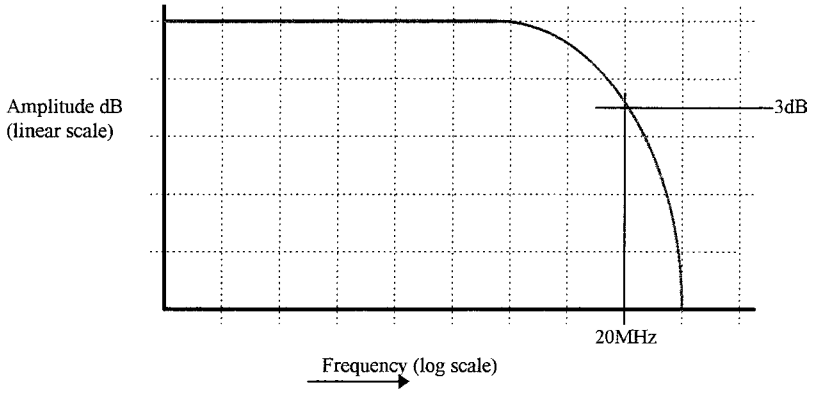


FIGURE 44 Frequency response curve for a dc-to-20-MHz oscilloscope.

by putting the time constant T of the oscilloscope circuitry equal to $1/\omega_{3\text{dB}}$ the displayed magnitude is 0.707 (70.7%) of the input. If T is made equal to $3/\omega_{3\text{dB}}$ the displayed magnitude is 0.95 of the input magnitude at $f_{3\text{dB}}/3$, i.e., at 6.7 MHz the response of our 20 MHz oscilloscope has fallen by 5%.

When measuring the rise time of a square wave or step function the rise time of the oscilloscope is of consequence. Since $t_{\text{rise}} = 2.2T$ and $f_{3\text{dB}} = 1/(2\pi T)$ it is evident that $f_{3\text{dB}} t_{\text{rise}} = 2.2/2\pi$ or 0.35. Hence for an oscilloscope with a -3 dB frequency of 20 MHz, the rise time is 17.5 nsec. The input impedance of an oscilloscope amplifier is, typically, 1 MΩ shunted by 20 pF; if this is connected to a signal source via a 1 m length of coaxial cable (typically the capacitance of 1 m of coaxial cable is 25 pF), the source will "see" an impedance of 1 MΩ shunted by 45 pF. At 1 MHz this measurement system load will be approximately 3.5 kΩ and in many situations will alter the operation of the circuit under observation. This means that the displayed waveform will be significantly different from that which was occurring before the measurement system was connected. In such a situation the problem of instrument loading can be very much reduced by using a passive probe between the oscilloscope and the signal source.

Figure 45 illustrates the passive R–C probe, which operates on the principle of the R–C divider (Section II.A) and provides a constant-voltage division for signals from dc to a practical upper limit (100 MHz or so). The correct adjustment of C_1 (the capacitance between the concentric

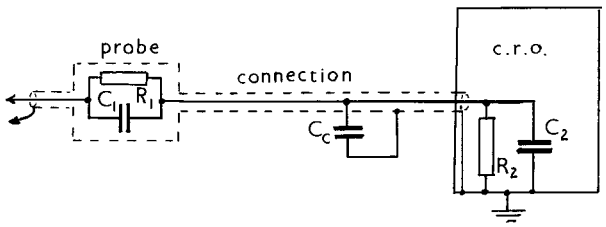


FIGURE 45 Passive R–C probe connected to a CRO.

core and the barrel of the probe) is set by observing the display of the oscilloscope's reference square wave, the correct value for C_1 giving the best square wave display. If the probe resistance R_1 is 9 MΩ and the cable capacitance 25 pF, C_1 will be 5 pF for a scope input impedance of 1 MΩ shunted by 20 pF. These values give a measuring system input impedance of 10 MΩ shunted by 4.5 pF, which at 1 MHz is 35.4 kΩ, or 10 times that without the probe. However, this decrease in loading is obtained at the expense of a reduction (to 1/10) in the amplitude of the displaced waveform. In general this can be compensated for by increasing the input amplifier gain of the CRO.

To monitor a small signal that is at a voltage above ground, e.g., displaying a current waveform as a voltage drop across a resistor, one must use either an oscilloscope fitted with a differential input or a conventional two-channel oscilloscope operated in its differential mode. This latter method of operation is achieved by use of the 'ADD' and 'INVERT' facilities available on most oscilloscopes. By selecting the switch positions that engage these two facilities, the displayed trace becomes the sum of the two input signals when the polarity of (normally) the channel 2 input has been reversed, i.e., the display is the input 1 signal minus the input 2 signal. While this method overcomes any isolation problems it results in a single trace display and so time differences between this displayed wave and the signal at any other point in the circuit under investigation cannot be determined. It is in these circumstances that one must use either an oscilloscope with a differential input or external circuitry to overcome the isolation problem.

The use of the X–Y mode of operation of an oscilloscope provides an alternative method of determining the time or phase difference between two waveforms. If the inputs to both channels are sine waves of the same frequency and one signal is applied to the vertical deflection plates and the other to the horizontal deflection plates, then, depending on the phase difference between the two signals, an elliptical display will be obtained such as in Fig. 46.

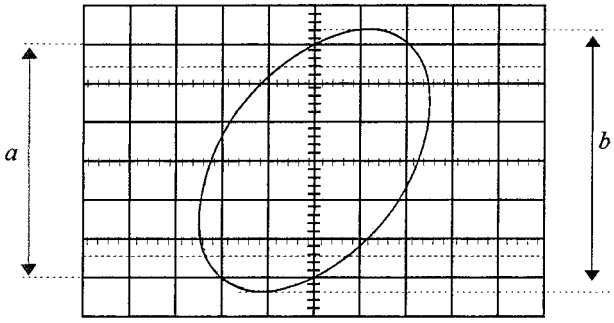


FIGURE 46 The determination of phase angle using an elliptical display.

Extracting the dimensions a and b from the display, it may be shown that the angle θ between the two signals is $\sin^{-1}(a/b)$ irrespective of the amplitude of either signal. For the ellipse in Fig. 46, $a = 6$ DIV and $b = 6.8$ DIV, and hence the phase angle is given by $\sin^{-1}(6/6.8) = 62^\circ$.

If the signals applied to the two inputs are of different frequency the ellipse will rotate at a speed dependent on the difference in their frequencies. This provides a very sensitive means by which one frequency may be adjusted to be the same as (or a simple multiple of) another.

B. Waveform Recording

While the continuous display of waveforms has many practical applications, the recording of waveforms in hard copy for analysis and comparison is equally necessary (particularly for nonperiodic waveforms). For slowly varying quantities, electromechanical recorders are usually used, while for frequencies of up to a few megahertz, recording oscilloscopes are used.

The earliest form of electric graphical recorder used a moving-coil milliammeter in which the pointer was modified to carry a pen that drew a trace on a clockwork-driven chart. This form of recorder had two shortcomings: first, since a moderate force was needed to drive the pen across the paper, an appreciable current through the deflecting coil was required, manifesting itself as a low input impedance and a moderate amount of power being drawn from the circuit under observation; and second, the deflection across the chart was in a circular arc rather than a straight line. Contemporary moving-coil pen recorders have overcome the loading problem by using an input amplifier and have linearized the deflection by using either a linkage system or electronic compensation in the input amplifier.

An alternative method of obtaining a linear movement across the chart is to use a potential difference or potentiometric system (Fig. 47). This provides a better accuracy than the moving-coil recorder, but in general can be used only to record signals whose frequency is below 2 Hz. It is very satisfactory for monitoring variables (e.g., transducer outputs) in a process plant and forms the basis of many X–Y plotters, an arrangement in which two potentiometric systems are used, one to provide the X deflection and the other the Y deflection.

To facilitate the display of nonperiodic waveforms that are too slow or too fast to be monitored on a conventional-general-purpose oscilloscope, various forms of storage have been used. For many years the most commonly used of these was the phosphor storage CRO, which had two stable states: written and unwritten. Once stored, the phosphor allowed waveforms to be displayed for up to several hours (unless erased by the operator). The fall in the cost of digital memory has provided a replacement to phosphor storage, namely, digital storage oscilloscopes. These have a superior performance to the phosphor storage scope in that an increased number of waveforms can be stored indefinitely, pre-trigger information can be displayed, waveforms can be transferred to a computer, hard copies of

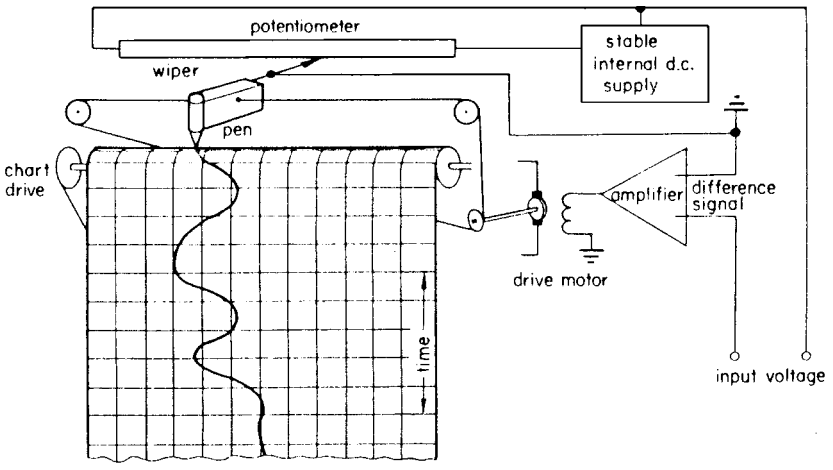


FIGURE 47 Components of a potentiometric pen recorder. [From Gregory, B. A. (1981). "An Introduction to Electrical Instrumentation and Measurement Systems," Macmillan Education, Hampshire, England.]

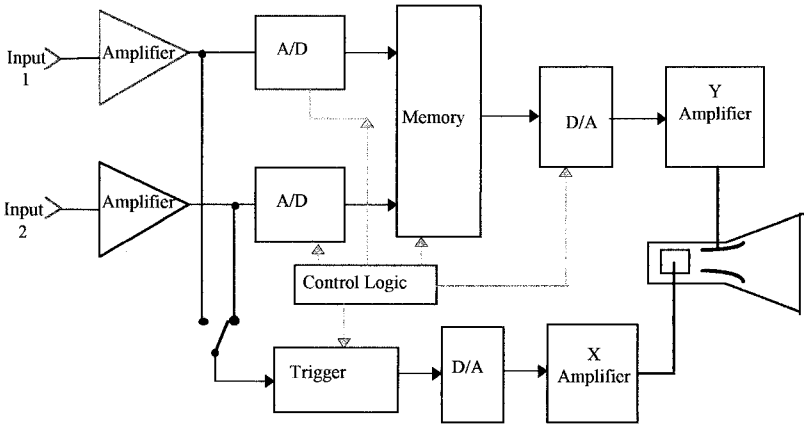


FIGURE 48 Schematic arrangement of a digital oscilloscope.

waveforms can be made, and mathematical processing of the waves may be made. In these devices (Fig. 48) the incoming signal is sampled at a rate controlled by the set time base, converted to a digital value, and stored in the memory. To view the waveform, the digital values are read from the memory, passed through a digital-to-analog (D–A), converter and displayed on a conventional CRT screen.

The developments in large-scale integration and liquid crystal displays have enabled the production of hand-held instruments that are a combination of a digital multimeter and a digital storage oscilloscope. Figure 49 illustrates one such instrument, which is ideally suited for use by many service engineers.

In many measurement situations it is desirable to record a number of variables simultaneously. Digital storage oscilloscopes generally have a two-channel capability; however, some can record and display four channels of analog input, while others can store up to 16 channels of digital signals for logic analysis purposes. The virtual instrument data acquisition modules that are used in conjunction with a PC can typically be used to monitor the outputs of 16 or 32 transducers whose output frequency is less than 2 Hz, for example, the output signals from thermocouples in a processing plant (see also virtual instruments, Section I.D).

FIGURE 49 Scopemeter® combined digital multimeter and oscilloscope. (Courtesy of The FLUKE Corporation.)

A further method of waveform recording that can be used is provided by instrumentation tape recorders, which use the same principles as the first magnetic recording of information that was performed by Valdemar Poulsen with his Telegraphon in Denmark in 1893. By 1935 the Germans had successfully developed a plastic tape, although in the United Kingdom and the United States, steel wire (or tape) was used until around 1950. Advances in materials have given the tape recorders of today a performance many times superior to that of the Telegraphon.

Tape recorders consist of three basic parts:

1. A recording head that creates a magnetic pattern in the magnetizable medium (a coating of fine magnetic oxide particles on a plastic tape)
2. The magnetic tape and tape transport system
3. The reproducing head that detects the magnetic pattern stored on the tape and translates it back into the original signal

Associated with each of these parts is substantial electronic and control circuitry to compensate for the characteristics of the heads, the inertia of moving parts, and so on.

In instrumentation recording, two techniques are in general use:

1. Direct recording, which has the wider bandwidth (50 Hz to 2 MHz), but suffers amplitude inconsistencies and “dropouts” due to tape coating inhomogenities.
2. Frequency-modulated recording, which overcomes the direct-recording problems, has a frequency range

from dc to 50 kHz, but requires more complex electronics than the direct- recording process.

C. Digital Frequency and Time Measurements

While the frequency of a signal can be established using an oscilloscope, the accuracy specification for most oscilloscope time bases gives a tolerance figure of ±5%. For many situations this is totally inadequate, as a more precise measurement is required. Fortunately, a sinusoidal wave can easily be converted into a pulse or square wave, which can then be counted by digital electronic circuits. With electronic counting techniques, the frequency of a signal can be determined with a very small tolerance.

The direct display of a frequency as a digital quantity appears to remove all problems in frequency measurement. However, an understanding of the operation of the processes used in a counter timer is of benefit since the manner in which the measurement is made has a direct bearing on the tolerance assignable to the reading. The digital counter timer may be considered to consist of a number of operational blocks that can be interconnected in different ways to perform various time-dependent measurements. The contents of these blocks of circuitry are as follows:

1. *Input circuits* which convert the incoming or unknown signal into logic levels compatible with the digital circuitry used within the instrument. Because of the wide range of signals that are likely to be encountered, both amplification and attenuation are provided as well as protection against accidental damage. The final stages of the input circuitry are a Schmitt trigger, which converts the input signal into logic pulses, as well as the trigger slope selection and level controls, which are essential for time-interval measurements.

2. The *reference oscillator* (sometimes referred to as the time base oscillator) is used to provide a known frequency or time interval. Some form of crystal oscillator is generally used to provide a reference frequency of 1 or 10 MHz. Although the stability of such oscillators is inherently good, it can be improved by temperature compensation circuits or, when used in the top-quality instruments, by being installed in a constant-temperature oven.

3. A string of *decade dividers* is necessary so that either the incoming signal or the reference frequency can be divided down. The number of decades of division used in a particular application is decided either by the operation of front panel switches or by the instrument's internal logic in the 'auto' mode of operation. The selection of decades in use is linked to the display so that the decimal point in the display is in the appropriate place.

By way of an example, seven decades of division would permit the production, from a 10-MHz reference, of time intervals of 1 μsec, 10 μsec, 100 μsec, 1 msec, 10 msec, 100 msec, and 1 sec as well as the 0.1 μsec directly from the oscillator.

4. A *main gate* (a dual-input logic gate) controls the passage of the pulses to the count circuits. Depending on the mode of operation being used, the conditioned input signal will either be counted (e.g., frequency measurement) or be used to control the counting of time-defined pulses (e.g., time interval measurement). The control circuits determine how the main gate will be used and when it will be enabled and disabled. This pulse-operated switch is arranged so that one command pulse opens the gate, i.e., starts the count, and the next command pulse closes the gate and thus stops the flow of pulses being counted. Figure 50 shows a schematic diagram with signal waveforms for a counting situation.

5. *The decimal counting unit and display* is the core of the instrument and consists of a number of counter

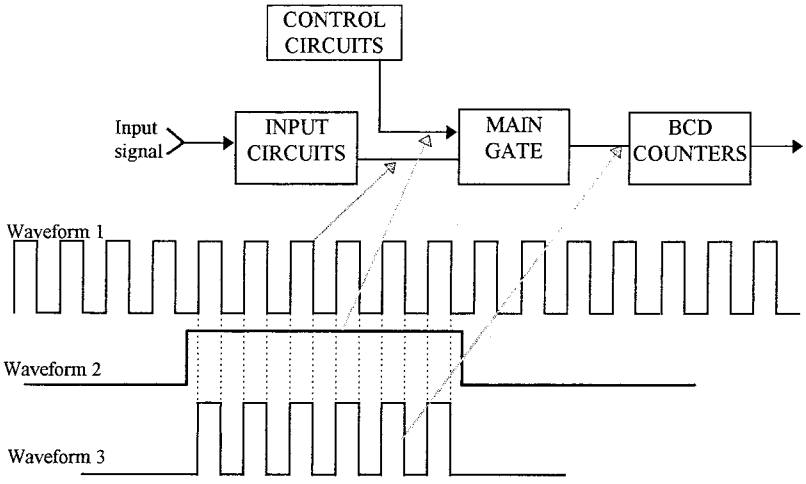


FIGURE 50 Schematic diagram and waveforms for the operation of the main gate.

decades in cascade. Each decade consists of a decade counter, memory, BCD-to-decimal decoder, numeral indicator driver, and numeral indicator. The actual display is commonly seven-segment LED or LCD numeral indicators arranged side by side to give a 6-, 8-, or even 10-digit display.

6. A set of *control circuits* is required to interconnect the above modules. These may be considered as a block of circuitry. The front panel of a counter timer has an array of switches and pushbuttons, some of which are used to select the mode of operation, for example, frequency, count, time interval, etc., and others that are used to control the duration of the measurement sample, the sensitivity, and so on. The switches that are used to select the mode of operation can, in some instruments, be used to activate the control logic so that either the counter samples and displays continuously the appropriate function of the incoming signal, or a single sample is made on demand and then 'held' as a display.

Most instruments also have an 'auto' or automatic ranging capability in their frequency mode of operation and in this condition their control circuits adjust the duration of the sample so that the display contains the maximum number of digits commensurate with a sample time of (usually) 1 sec.

The measurement functions available for a particular counter timer are dependent on its complexity (and cost). The simplest function available on any counter timer is the count or totalling mode of operation, in which the input signal is routed via the input circuitry to the main gate. External command pulses are then used to open (and then close) the gate for the duration of the count.

To obtain a direct display of the frequency of a signal simply requires that a count of cycles of the input wave be made over a known time interval. Figure 51 shows the interconnection of the circuit blocks that provide these conditions. The count of the unknown signal is made over 1000 circles of the reference frequency. While such a process is very satisfactory for frequencies above 10 kHz, to obtain a good-resolution (six-digit) display of a low frequency, say 100 Hz, would require an undesirably long measurement time: 100 sec would be required to provide a display of 100 Hz.

To overcome the problem of poor resolution and hence large tolerances when measuring frequencies lower than 1 kHz, the circuit blocks can be rearranged as in Fig. 52 so that a count of reference oscillator pulses is made in one or a number of cycles of the unknown frequency. Then, for a 100-Hz signal divided by 100, a count of the reference oscillator (1 MHz, say) is made for approximately 1 sec (i.e., a count of 1,000,000), giving a much improved resolution and smaller measurement tolerance. The disadvantage is that it is then necessary to calculate the frequency of the input signal by determining the reciprocal of the displayed value.

To overcome the need for the operator to perform the reciprocal calculation, contemporary counter timers that include a microprocessor for control purposes use its arithmetic processing capability to perform the reciprocal calculation and display the frequency value. Instruments with such capabilities are variously referred to as reciprocal or calculating counters. The general operation of a reciprocal counter is essentially the same as that of a conventional counter, except that when operating in the

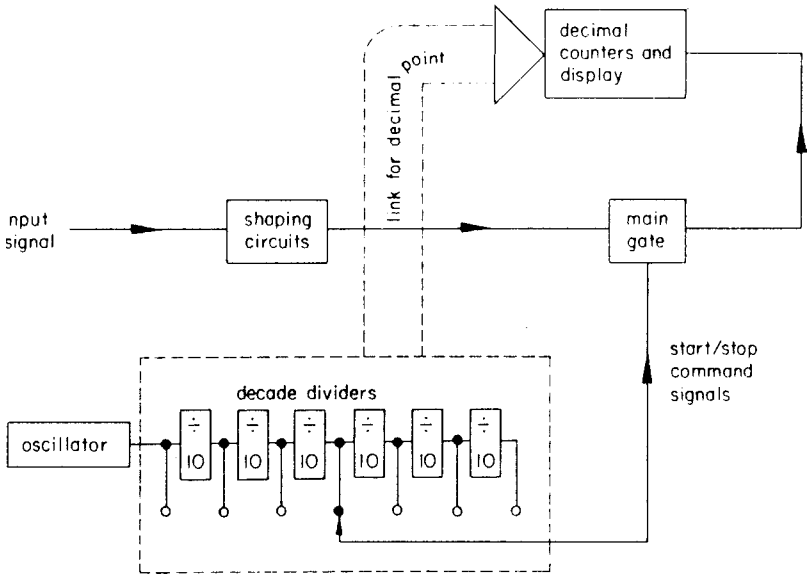


FIGURE 51 Schematic diagram of connections for frequency measurement. [From Gregory, B. A. (1981). "An Introduction to Electrical Instrumentation and Measurement Systems," Macmillan Education, Hampshire, England.]

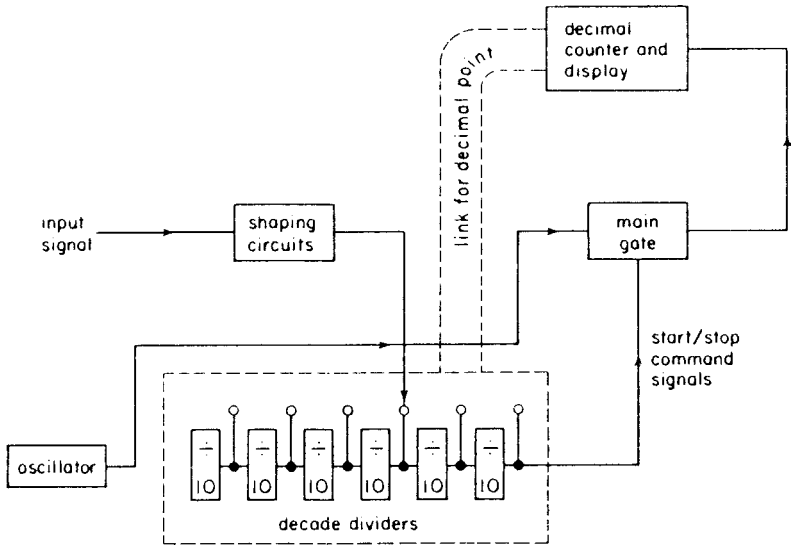


FIGURE 52 Schematic diagram of connections for multiple period measurement. [From Gregory, B. A. (1981). "An Introduction to Electrical Instrumentation and Measurement Systems," Macmillan Education, Hampshire, England.]

frequency mode and measuring frequencies less than the reference oscillator frequency, a reciprocal counter makes simultaneous counts in separate registers (for an operator-controlled gate time) of clock pulses (time) and cycles of the unknown frequency (events). On completion of the gate time, the processor computes the frequency from the event and time counts. Figure 53 is a schematic of the circuit modules in a reciprocal counter. The duration of the adjustable gate time is in whole cycles of the unknown frequency, it being possible to measure and display the duration of the gate time by suitable setting of the front panel controls.

In the time-interval mode of operation (Fig. 54), a clock unit of suitable duration, say 1 μsec or 1 msec, can be selected and used to measure the time interval between positive- or negative-going slopes at the zero crossing or (in a sophisticated instrument) at an operator-selectable level. The points between which it is possible to measure a time interval may be on a single input wave or on two separate input signals. When making small-time-interval measurements on separate signals care must be taken to match the connection lengths, or an appreciable error can be introduced. This time-interval facility is very useful when measurements of mark-space ratios or phase displacements have to be made.

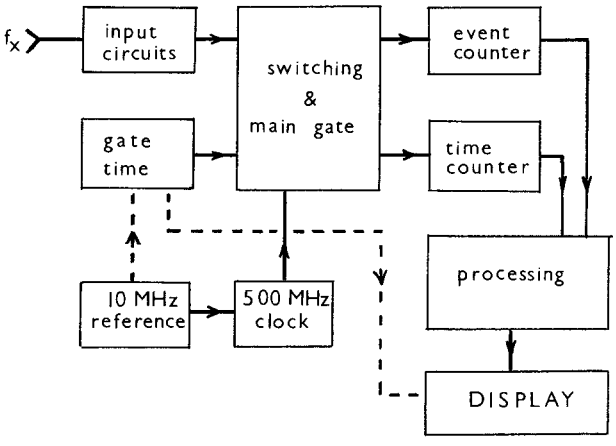


FIGURE 53 Schematic diagram of connections for frequency measurement in a reciprocal counter timer.

D. Spectrum Analysis

It is conveniently assumed that alternating signals are single-frequency sinusoidal waveforms. Although this is fortunately an acceptable assumption for many situations, the presence of harmonics in signal waveforms cannot always be ignored, and their magnitudes must be established. Figure 55 (see page 33) presents a waveform that consists of a fundamental, a third harmonic, and a fifth harmonic as a function of time (Fig. 55a) and frequency (Fig. 56b) and as a three-dimensional form in an attempt to link the two (Fig. 55c).

The conventional CRO has a horizontal scale that is a simple function of time. In many measurement situations this is unsatisfactory because what is really required is the analysis and display of the signal in the frequency domain. The spectrum analyzer provides the visual display of amplitude against frequency. It was originally developed for the analysis of the components of radiofrequency signals and shows in a single display the Fourier components of a given waveform (Fig. 55a). Many contemporary spectrum analyzers have the computing power to perform

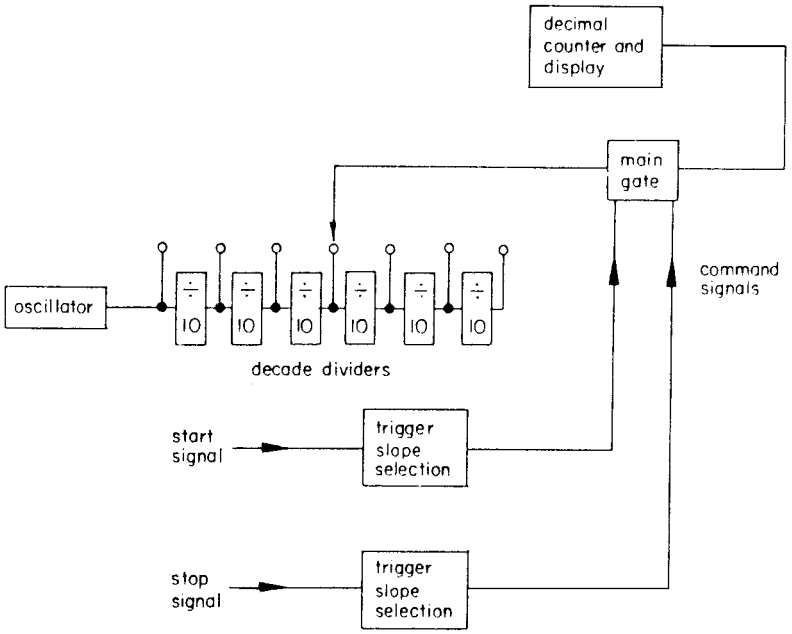


FIGURE 54 Schematic diagram of connections for time interval measurement. [From Gregory, B. A. (1981). "An Introduction to Electrical Instrumentation and Measurement Systems," Macmillan Education, Hampshire, England.]

the Fourier analysis and other measurement functions on the applied waveform. Some virtual instrument packages (Fig. 11) have spectrum analysis capabilities as well as multimeter and oscilloscope facilities.

The applications of the spectrum analyzer are many. For example, acoustic noise and vibration levels are of major concern to manufacturers and users of mechanical vehicles; by using appropriate transducers, electrical signals derived from the vibrations can be examined to assist in locating the source. In the fields of electronics and communications, spectrum analyzers can be used to provide performance information on carrier-wave purity, modulation, frequency response, and electrical noise, in measurements concerned with (1) identifying signals resulting from nonlinear amplification, filtering, or mixing, (2) determining the purity of signals, (3) measuring and displaying frequency and modulation characteristics, (4) determining the frequency response of a network, and (5) analyzing of electromagnetic compatibility performance.

IV. NONELECTRICAL QUANTITIES

Measurements of many physical quantities such as length (position), pressure, force, temperature, and rainfall are performed using electrical and electronic instruments or measuring systems. In an arrangement that performs such measurements, it is necessary to have a device that converts the physical quantity into related values of an electrical parameter. These devices are termed *transducers*. While this section is concerned only with the electrical measurement aspects of the devices, since it necessary to ensure that the record or display is a true representation of the measurand, it is worth briefly considering the principal methods by which this electrical output is achieved.

1. Resistance Change

(a) In *potentiometric* devices (Fig. 56) the variation in displacement of the measurand causes the wiper to move along the resistor and provides an output voltage that is proportional to displacement.

(b) Many transducers incorporate *strain gauges* that can also be used in their own right for the determination of strains in members of a structure. The strain gauge is based on the principle that if a conductor is stretched, its length increases at the expense of its cross-sectional area. Both these dimensional changes result in an increase in its resistance R since $R = \rho l / A$, where ρ is the resistivity of the conductor, l is its length, and A is its cross-sectional area. Since long lengths of resistance wire would be inconvenient to use, the conductor is formed into a grid bonded to an insulating material as in Fig. 57. The complete gauge is glued to the member of a structure being investigated.

(c) Some transducers use the *temperature coefficient of resistance* to provide the parameter that is sensitive to the measurand. The most notable of these is the platinum resistance thermometer, which has proved to be very

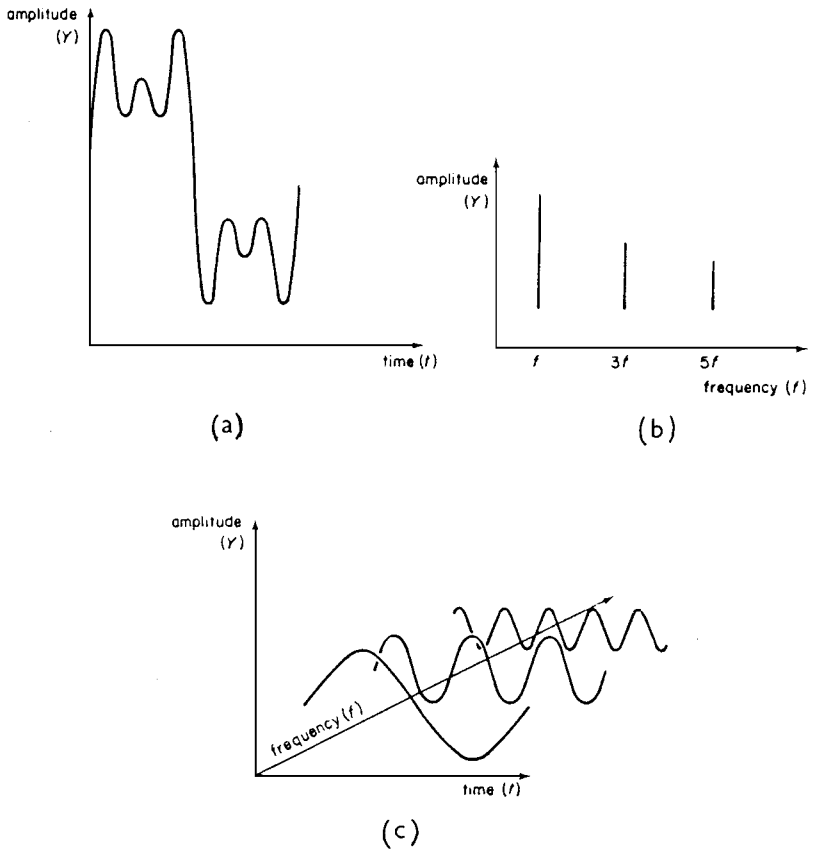


FIGURE 55 Presentation of multifrequency waveform on time and frequency axes. [From Gregory, B. A. (1981). "An Introduction to Electrical Instrumentation and Measurement Systems," Macmillan Education, Hampshire, England.]

reliable for industrial applications, its resistance being monitored by one of the techniques outlined in Section II.D. When used with specially developed precision bridge circuits, the platinum resistance thermometer is used as a reference standard for temperature measurement.

2. Reactance Change

(a) A change in the *inductance* value of an inductor may be achieved by varying the reluctance of its core by changing the magnitude of an air gap in it or by moving a slug of high-permeability material into it. Both of these techniques have been used to create variable-inductance transducers. A variation of the latter arrangement that is extensively used is the linearly variable differential transformer (LVDT), in which the coupling between a central winding to adjacent outer windings is varied by the movement of a high-permeability core (Fig. 58).

(b) In its simplest form a capacitor consists of a pair of plates separated by a dielectric (air). Variation in *capacitance* can be achieved either by changing the separation of the plates or by changing the overlap area of the plates.

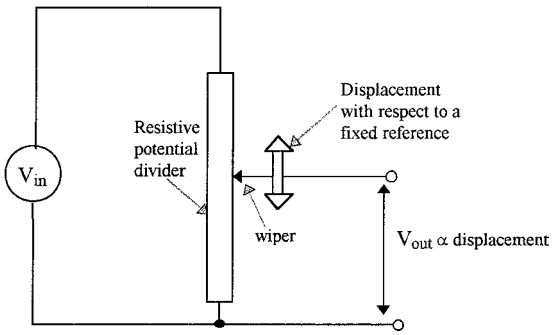


FIGURE 56 Principle of the potentiometric transducer.

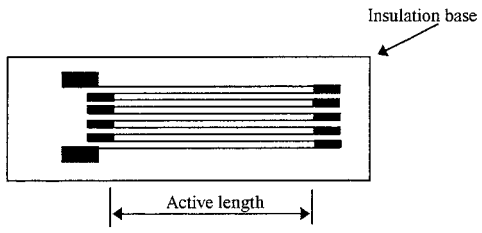


FIGURE 57 Simple strain gauge.

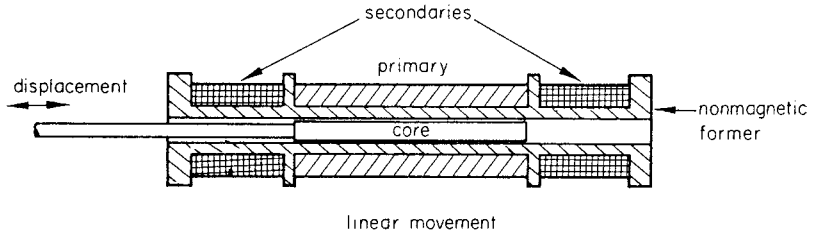


FIGURE 58 A linearly variable differential transformer. [From Gregory, B. A. (1981). "An Introduction to Electrical Instrumentation and Measurement Systems," Macmillan Education, Hampshire, England.]

3. Self-Generating

(a) *Electromagnetic*: The movement of a magnet within a coil will induce a voltage in the coil. This property is utilized in vibration transducers, which measures linear movement, and in which the magnet is vibrated within the coil; in tachogenerators, in which a magnet is rotated in a coil system so a voltage proportional to the speed of rotation is generated; and in toothed-rotor pulse-generating transducers, which produce a voltage pulse each time a tooth passes a coil wound around a magnet, with the frequency of the pulse waveform directly related to the speed of rotation and the number of teeth on the rotor.

(b) *Thermoelectric*: Probably the most widely used (and least expensive) of all transducers is the thermocouple, which consists of a pair of conductors that are made from different materials, for example, copper and constantan. Commercially one of the most widely used conductor pairs is the nickel/chromium–nickel/aluminium (*chromel–alumel*) combination, which is obtainable with plastic insulation for laboratory use or in a mineral-insulated metallic sheath for industrial applications. Perhaps its major drawback is that the output signal is not directly proportional to the temperature difference between hot and cold junctions, which makes it necessary to insert some linearization by using either (originally) electronic hardware or (nowadays) computer software between the electrical output of the thermocouple and the display of temperature.

(c) *Piezoelectric*: If crystals of quartzlike materials are squeezed, an electric charge appears on the crystal surface ("*piezo*," from the Greek "*to squeeze*"). This charge can be measured via a charge amplifier (one with a high input impedance, say 100 MΩ) and a recording instrument. Since the charge dissipates through the measuring system these piezoelectric transducers should be used only in a dynamic situation. Their main application is for the measurement of acceleration, impact, and related conditions.

A. Transducer Outputs

In addition to the self-generating devices referred to above, many transducers are arranged so that their output signal is a voltage whose amplitude is directly proportional to the magnitude of the measurand. The advantage of such an arrangement is that a display in engineering units is easily obtained with simple scaling techniques. The convenience of this form of output signal has resulted in many transducer types that require ac energization; for example, some LVDTs are manufactured with incorporated electronics that convert a dc power supply to ac for operation of the transducer, its output being rectified to produce a direct voltage output with an amplitude that is proportional to the displacement in the measurand. While such incorporated electronics can be arranged to provide an output of satisfactory amplitude, difficulties arise with some transducers due to the small size of their output signal and effective output resistance. Typical values for a transducer that incorporates a strain gauge bridge are 10 mV maximum and 120 Ω minimum, respectively. An exception to this is found with transducers that use the potentiometer principle since these devices may have an output amplitude that (typically) can rise to 10 V, but these generally have a high output resistance and a limited operational life.

B. Signal Transmission

In many practical situations the transducer is situated in a location remote from the monitoring or measuring instrument. The effective output resistance of many transducers has an appreciable magnitude, whereas the amplitude of the output signal may be small. These conditions require the use of sensitive, high-input-impedance measuring instruments and result in a situation almost guaranteed to provide interference problems. In consequence, it is usually necessary to use screened connections between the transducer and the monitor. If the signal is very small, a guarded instrument/measuring circuit (Fig. 21) will be needed to overcome the combined effects of common mode voltages and lead resistances. The alternative to elaborate guarding is the provision of signal processing at the transducer, of which there are various forms:

1. Voltage amplification, so that the transmitted signals are large compared with the interference levels.

2. Conversion to current. The "current-loop" method of signal transmission converts the transducer output to a current level that varies between 4 and 20 mA. The signal reverts to a voltage at the instrument end as a voltage drop across a fixed resistor. If the current falls below 4 mA, a fault condition is considered to exist.

3. Conversion to light signals, which are transmitted along optic fibers. This is a fairly expensive undertaking but is unaffected by electrical interference and is finding favor in some hazardous environments.

4. Digitizing at the transducer. Since digital signals are largely unaffected by lead resistance and low-frequency electrical interference, there are advantages in transmitting digital rather than analog signals. However, the only true digital transducer is the position encoder; consequently, the advantages of digital transmission can only be obtained by converting the output of an analog transducer into a digital signal. This has become feasible because of the miniaturization of electronic components, and transducers with built-in digitizers can be manufactured that not only perform the A–D function but also provide processing and memory capabilities.

5. Conversion to radiofrequencies. When transducers are located in remote situations, such as for weather monitoring on a mountain, the transmission of the data can be performed using a radio link. While such a process is moderately common in situations such as the chosen example, it is rarely satisfactory in an industrial environment due to the noise levels.

A transmission problem peculiar to thermocouples is due to the use of a material different from that of the thermocouple being used in the transmission path. The probable result is the creation of a "reference" junction at a point of variable temperature and the introduction of a seemingly random error into the temperature monitoring. The problem is usually avoided only by running the thermocouple conductors right to the monitoring instrument, which must incorporate automatic reference junction compensation.

C. Instrument Loading

The effects of instrument loading in monitoring the output from transducers must always be considered since most types of transducer appear to the measuring arrangement as a signal source with a moderate to high output impedance. In consequence, unless the instrument's input impedance is sufficiently high, a considerable insertion error will occur (Section I). If the measurand is oscillatory in nature, the frequency response of the measuring system must be adequate, it being remembered that if this is specified as a bandwidth, the bandwidth of the system is

$$B_s = \frac{B_1 \times B_2}{\sqrt{B_1^2 + B_2^2}} \text{ Hz},$$

where B_1 and B_2 are, respectively, the bandwidth frequencies of the transducer and the measuring instrument. The bandwidth frequency is that at which the response of the device is reduced by 3 dB (or 30%), that is, the output signal V_0 is 0.707 of the mid-frequency value [and the output power is half of the mid-frequency value, from the following expression: response or gain (in dB) $= 10 \log_{10}(V_0^2 / V_{\text{ref}}^2)$]. To obtain a frequency at which the system response is still within 1% of the specified value requires the 3 dB frequency to be divided by five (see Section III.B).

When transducers are operated in remote locations, to conserve battery power there is an increasing trend to connect the electrical supply to the transducer only while readings are being made. If such a form of operation is adopted, sufficient time must be allowed so that the transducer output has settled to its correct value before the reading of the measurand is made.

V. AUTOMATED MEASUREMENTS

A. Recording and Monitoring Methods

The early efforts to automate measurements, that is, to obtain a record of readings without the presence of an operator, resulted in modifying a moving-coil meter to produce a graphical recording of a single variable. The contemporary versions, single and multichannel, of such instruments (described in Section III) are widely used for recording process variables.

The development of the digital voltmeter in the early 1960s provided an instrument that could be connected through a switch to sample a number of unknown signals. Because of the availability of a digital output from the instrument, it became possible to provide a printed record of the sampled measurands. By adding a time clock, we arrive at the data-logger (Fig. 59). The contemporary versions of the data-logger are many and various, from small devices that have limited digital memory (one or two input channels, suitable for use in remote locations), to versatile multichannel (up to 1000) programmable loggers and some virtual instrument (data acquisition) systems, suitable for complex research investigations or permanent monitoring of a process plant.

One of the problems of automated measurement is the ease with which one can accumulate vast quantities of data. To avoid this in permanent logger installations, the operation is programmed so that although the measurands are continuously scanned, a printout is provided only if

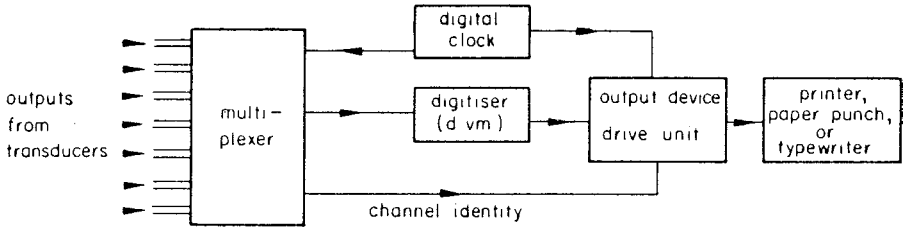


FIGURE 59 Block diagram of a basic data logger.

a preset limit is exceeded or when a set of readings is required for record purposes.

An alternative arrangement that is provided in some monitoring instruments is the use of a memory to store a limited amount of information. This store is continuously updated, and when a fault condition occurs, the information prior to the event can be printed out.

The trend in recent years to incorporate a microprocessor and memory within a digital multimeter has provided programmable and logging facilities such as displaying a reading multiplied by a set constant or the deviation of a reading from a set level, storing 100 readings, and being able to recall the maximum, minimum, and average values.

B. Bus-Connected Measurement Systems

To enable instruments to be coupled together and controlled in a preprogrammed manner, an internationally accepted bus arrangement has been adopted. It is known variously as the IEEE 488 interface bus, the HPIB (Hewlett Packard interface bus), and the GPIB (general-purpose interface bus). It evolved from the HPIB that was originally formulated in the late 1960s.

This widely used bus has specified mechanical and operational protocol arrangements that enable instruments of different manufacture to be linked to a programmed controller (computer). The bus consists of eight parallel data lines, three handshake or control lines, five management lines, and seven ground or return lines (one for data, three for the handshake circuits, and three for particular management lines). The IEEE bus connectors are of a stackable type, having 24 pins (1 is connected to the screen around the 23 connections). Up to 15 devices can be interconnected in a single system, provided a 20-m length of bus is not exceeded. However, if a modem is used, a system can be extended over long distances via a telephone link, although this seriously affects the operating speed.

Within a system a controller commands devices designated as *talkers*, which can only put data onto the bus (e.g., some measuring instruments), *listeners*, which can only receive data and instructions from the bus (e.g., some printers and switch units), and *talkers and listeners*, which can both receive and transmit data (e.g., multifunction meters, which can have range and function remotely changed).

The controller may be a microcomputer specially designed as a bus controller, a personal computer fitted with an interface to the bus, or a technical computer if speed and complexity of the measuring system, together with other computer requirements, justify the expense. The programming language used with most controllers is user-friendly, frequently being a modified form of BASIC.

The devices connected via an IEEE bus system may be arranged in a linear fashion, as a star, or as a combination of line and star. So that a particular device will know when it is required to perform a task, every device (talker or listener) must have a bus address that is unique within the system. To facilitate this, every IEEE bus-compatible device has a binary bus address that is switch-settable to values between 0 and 30. These switches are usually located on the back panel of a device but may be mounted internally; occasionally they are software controlled and accessible via the front panel key pad. For this latter arrangement, the address is stored in nonvolatile memory.

The benefits of the IEEE bus in forming automated or controlled measurements are immense, for the bus makes it possible to take large numbers of readings, process figures, and then tabulate results or plot graphs in a short time. The bus has found a very large number of uses in research, development, and teaching laboratories and for testing in a production situation.

For situations where the operation of the IEEE bus is not fast enough the VXIbus has been developed around the VMEbus architecture. This enables high-speed data rates of 40 MB/sec along with the necessary communication protocols to provide the means of building instrumentation systems for high throughputs. In addition, the design of its architecture allows for the integration of VXIbus products into traditional IEEE test systems as well as standalone systems.

C. Automated Test Equipment

The complexity of electronic equipment is such that it is impractical manually to test it thoroughly, and the

only way in which a manufacturer can make performance checks or tests on products is by computerized testing. This has led to the rapid growth of the automated test equipment (ATE) industry. The electronics industry has always needed to test its products, and this requirement was met by the in-house manufacture of equipment for testing sections or subassemblies of a particular system.

The justification for automated testing is economic. A company must provide its customers with reliable working products or go out of business, hence it must test its manufactured equipment. In devising a testing strategy, a company needs to analyze its product range, scale of production, anticipated new products, and the types of fault that occur during manufacturing processes.

The forms of testing available to a manufacturer are the following:

1. Component test. Checks individual components prior to assembly, including checking printed circuit boards for open and short (between track) circuits.
2. In-circuit test. As in item 1, but after the components have been inserted and soldered to the printed circuit board. Frequently this type of test is preceded by a visual inspection to ascertain the absence of any components and so forth.
3. Functional test. Exercises the board in a manner that simulates its use in the final product.
4. Combination test. Combines the in-circuit and function test situations and overcomes the fault location/diagnosis problems of the functional testers.
5. Subassembly/system test. Function tests the subassemblies or the complete system.

Since it is generally accepted that the further along a production process a fault is found the greater the cost, the arguments for component testing are strong. However, the actual number of faulty components at purchase is quite small. The greater problem is incorrect insertion of components, such as diodes the wrong way round or wrong-value resistors inserted. As a result of this, many companies omit component tests in favor of in-circuit testing. The problem with in-circuit testing is that it requires "bed-of-nails" test fixtures, one for each printed circuit board, to make contact at appropriate points within a circuit, both to connect to the component under test and to guard out the effects of parallel paths (see Section II.D). A further disadvantage is that components are unlikely to be checked at their intended operational speed, so although they appear satisfactory at a low test frequency, they may not operate correctly at operational frequency. The advantage of in-circuit testing is that every faulty component can be located in a single test.

Functional testing overcomes the operational condition shortcomings of in-circuit testing and the need for the expensive test fixture, but in general it can identify only one fault at a time. This has led to the provision by some ATE manufacturers of combination testers.

To enhance the capabilities of ATE systems, many have an IEEE bus interface. This allows for complex signal patterns to be applied to the equipment under test and for specialized instruments to be used to make measurements on its output.

SEE ALSO THE FOLLOWING ARTICLES

ACOUSTICAL MEASUREMENT • ANALOG-SIGNAL ELECTRONIC CIRCUITS • CIRCUIT THEORY • CRITICAL DATA IN PHYSICS AND CHEMISTRY • ELECTROMAGNETIC COMPATIBILITY • SIGNAL PROCESSING, DIGITAL • SUPERCONDUCTIVITY

BIBLIOGRAPHY

Fluke Corp. (1994). "ABC of Oscilloscopes," Fluke Corporation, Everett, WA.

Bolton, W. (1996). "Measurement and Instrumentation Systems," Newness, Oxford.

Gregory, B. A. (1981). "An Introduction to Electrical Instrumentation and Measurement Systems," Macmillan, Basingstoke, UK.

Gregory, B. A. (2000). "An Introduction to the Use of an Oscilloscope," TekniCAL, Feedback Ltd., Crowborough, UK.

Kularatna, N. (1996). "Modern Electronic Test and Measuring Instruments," Institution of Electrical, Engineers, London.

Morris, A. S. (1996). "The Essence of Measurement," Prentice Hall, London.

Usher, M. J., and Keating, D. A. (1996). "Sensors and Transducers," (2nd ed.), Macmillan, Basingstoke, UK.

Flow Visualization

Wolfgang Merzkirch
Universität Essen

GLOSSARY

Image processing Digitization of a recorded flow picture and subsequent evaluation of the pattern by a computer to determine quantitative data.
Laser-induced fluorescence Visualization by tracer material that emits fluorescent radiation upon excitation by laser light.
Line-of-sight method Visualization by transmitting a light wave through the fluid flow.
Tomography Computer-aided reconstruction of the three-dimensional, refractive-index distribution in a flow to which a line-of-sight method has been applied in various viewing directions (projections).
Tracer particles Foreign particles with which a fluid flow is seeded for the purpose of flow visualization by light scattering.
Whole-field method Diagnostic method providing the information for a whole field of view (photograph) at a specific instant of time.

THE METHODS OF FLOW VISUALIZATION are diagnostic tools for surveying and measuring the flow of liquids and gases. They make visible the motion of a fluid that is normally invisible because of its transparency. By applying one of the methods of flow visualization, a flow picture can be directly observed or recorded with a camera. The information in the picture is available for a whole field, that is, the field of view, and for a specific instant of time (*whole-field method*). The information can be either qualitative, thus allowing for interpreting the mechanical and physical processes involved in the development of the flow, or quantitative, so that measurements of certain properties of the flow field (velocity, density, and so forth) can be performed. The techniques of flow visualization, which are used in science and industry, can be classified according to three basic principles: light scattering from tracer particles; optical methods relying on refractive index changes in the flowing fluid; and interaction of the fluid flow with a solid surface.

I. PRINCIPLES OF FLOW VISUALIZATION TECHNIQUES

The methods of visualizing a flowing fluid are based on the interaction of the flow with light. A light wave incident into the flow field (*illumination*) may interact with the fluid in two different ways: (1) light can be scattered from the fluid molecules or from the tracer particles with which the fluid is seeded; and (2) the properties of the light

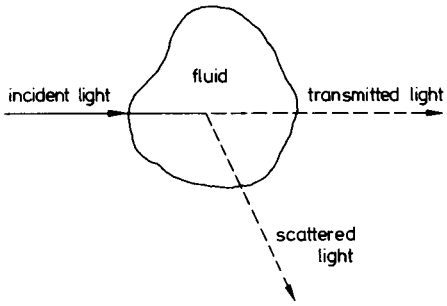


FIGURE 1 Light incident into a fluid flow is scattered from the fluid molecules or from tracer particles in various directions. The light wave transmitted through the transparent fluid is altered in comparison to the incident light, if the flow field exhibits changes of the fluid's refractive index.

wave can be changed, because of a certain optical behavior of the fluid, so that the light wave transmitted through the flow is different from the incident light (Fig. 1). The visualization methods based on these two interaction processes are totally different in nature and apply to different flow situations.

Since the light scattered from the fluid molecules (*Rayleigh scattering*) is extremely weak, the flow is seeded with small tracer particles (dust, smoke, dye, and so forth), and the more intense radiation scattered from these tracers is observed instead. It is thereby assumed that the motion of the tracer is identical with the motion of the fluid, an assumption that does not always hold, for example, in nonstationary flows. The scattered light carries information on the state of the flow at the position of the tracer particle, that is, the recorded information is local. For example, if the light in Fig. 1 is incident in form of a thin light sheet being normal to the plane of the figure, an observer could receive and record information on the state of the flow (e.g., the velocity distribution) in the respective illuminated plane.

The signal-to-noise ratio in this type of flow visualization can be improved if the tracer does not just rescatter the incident light but emits its own, characteristic radiation (*inelastic scattering*). This principle is realized by fluorescent tracers (e.g., iodine), which may emit bright fluorescing light once the fluorescence is induced by an incident radiation with the appropriate wavelength (*laser-induced fluorescence*).

An optical property of a fluid that may change the state of the transmitted light wave (Fig. 1) is its index of refraction. This index is related to the fluid density, so that flows with varying density, temperature, or concentration may affect the transmitted light. Two effects are used for such an optical flow visualization: the deflection of the light beams from their original direction (*refraction*) and the change of the phase distribution of the wave. The latter is measured by means of optical interferometry; shadowgraph and schlieren methods are sensitive to the refractive light deflection, and they are used for qualitative optical flow visualization.

The information on the state of the flow or the density field is integrated along the path of the transmitted light, that is, the information is not local as in the case of the techniques using light scattering. For the purpose of a quantitative information on the three-dimensional flow field, it is necessary to desintegrate (invert) the data, which is recorded in two-dimensional (plane) form, for example, on a photograph. This requires the application of methods known as computer tomography.

The result of a flow visualization experiment is a flow picture that may be recorded with a camera (for example, a photographic camera, movie camera, or video camera). The information on the state of flow is available in the recording plane (x–y plane) for a specific instant of time, t_i; or for a number of discrete instants of time in the case of kinematic recording. This makes the information different from that obtained by probe measurements (e.g., hot wire anemometer, laser-Doppler velocimeter), where the data is measured only at one specific location (coordinate x, y, z), but as a continuous function of time t.

For the purpose of generating quantitative data on the flow, the recorded visual pattern must be identified, evaluated, and interpreted. In order to objectify the evaluation and make it independent of a viewer's interpretation, the pattern can be recognized by electro-optical devices and transformed into digital data, which are then processed by a computer (*image processing*).

II. FLOW VISUALIZATION BY TRACER MATERIAL

That a flow becomes visible from foreign particles that are floating on a free water surface or suspended in the fluid is a fact of daily experience. This crude approach has been refined for laboratory experiments. The methods of flow visualization by adding a tracer material to the flow are not real science but art that concerns the selection of the appropriate tracers, their concentration in the fluid, and the systems for illumination and recording.

The trace material, after being released, is swept along with the flow. If one does not resolve the motion of single particles, qualitative information on the flow structure (streamlines, vortices, separated flow regimes) becomes available from the observed pattern. The identification of the motion of individual tracers provides quantitative information on the flow velocity, provided that there is no velocity deficit between the tracer and the fluid. Only in the case of fluorescent (or phosphorescent) tracers is it

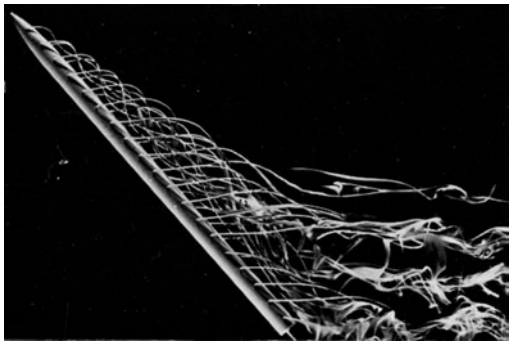


FIGURE 2 Visualization of the swirling flow behind an inclined cylinder by dye lines in water. Original dye lines are alternatingly red and blue. (Courtesy of Dr. H. Werlé, ONERA, Châtillon, France.)

possible to deduce data on quantities other than velocity (density, temperature).

Besides some general properties that any seed material for flow visualization should have (e.g., nontoxic, noncorrosive), there are mainly three conditions the tracers should meet; they are neutral buoyancy, high stability against mixing, and good visibility. The first requirement is almost impossible to meet for air flows. Smoke or oil mist are the most common trace materials in air, with the particle size of these tracers being so small ($<1\ \mu$m) that their settling velocity is minimized. A number of neutrally buoyant dyes are known for the visualization of water flows, the colors introducing an additional component of information (Fig. 2).

Special arrangements for illumination and recording as well as timing are necessary if the goal is to measure the velocity of individual tracer particles. A time exposure is a possible way for visualizing the instantaneous velocity distribution in a whole field (plane) of the flow. Each particle appears in the form of a streak whose length is a measure of the velocity vector in the plane (Fig. 3). An alternate way is to take a double exposure produced by two very short light pulses with a definite time interval between the two pulses. An optical or numerical Fourier analysis of the field of particle double images provides the distribution of the velocity vectors ("particle image velocimetry": PIV). The plane section in the flow is realized by expanding a thin laser light beam in one plane by means of a cylindrical lens, so that all tracer particles in this plane light sheet are illuminated. The velocity component normal to the plane

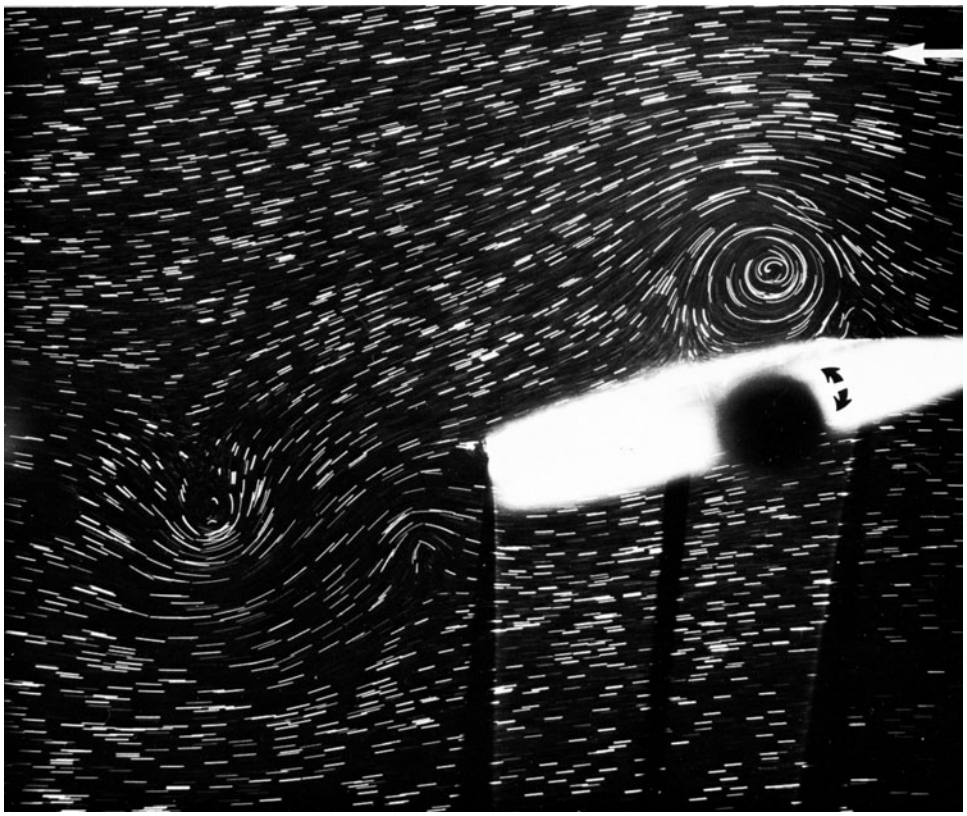


FIGURE 3 Time exposure of the water flow around a pitching airfoil shape. The water is seeded with tracer particles that appear as streaks of finite length. The pattern of the streaks is a measure of the distribution of the velocity vector in the illuminated plane. (Courtesy of Dr. M. Coutanceau, Université de Poitiers, France.)

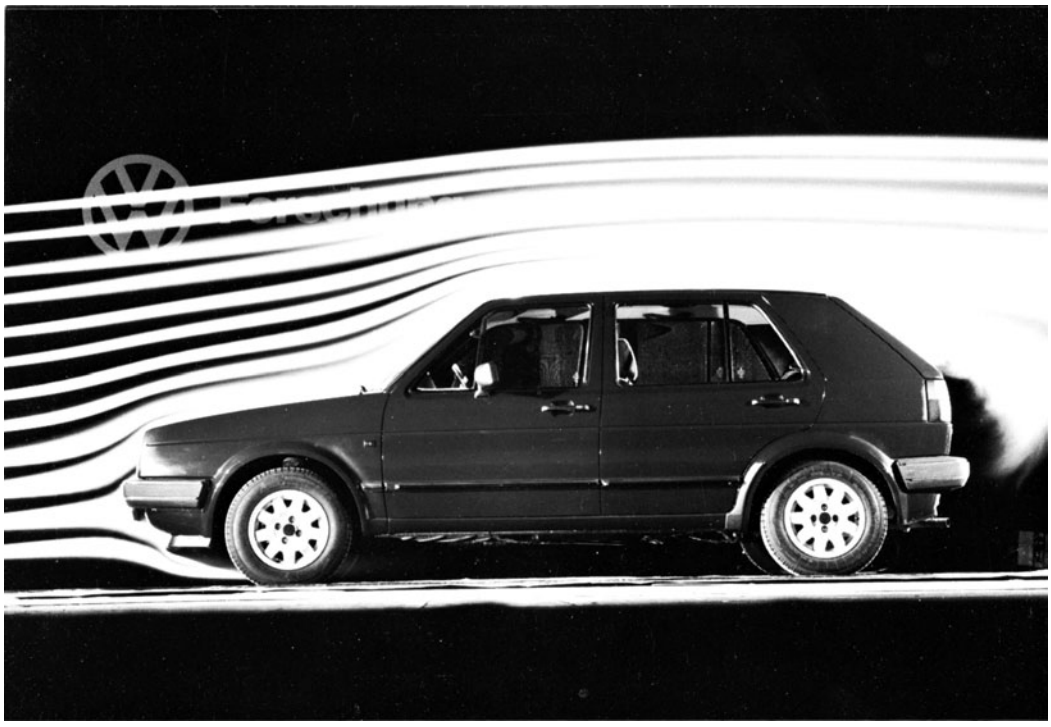


FIGURE 4 Smoke lines around a car in a full-scale wind tunnel. (Courtesy of Volkswagenwerk AG, Wolfsburg, Federal Republic of Germany.)

is not recovered. Flow visualization by tracer materials is a standard technique in wind tunnels, water tunnels, other flow facilities, and field studies. A typical application is the study of the flow around car bodies with the aim of improving the aerodynamic characteristics of the shape (Fig. 4).

III. VISUALIZATION BY REFRACTIVE INDEX CHANGES

The refractive index of a (transparent) fluid is a function of the fluid density. The relationship is exactly described by the Clausius-Mosotti equation; for gases, this equation reduces to a simple, linear relationship between the refractive index, n, and the gas density, ρ, known as the Gladstone-Dale formula. Therefore, refractive index variations occur in a fluid flow in which the density changes, for example, because of compressibility (high-speed aerodynamics or gas dynamics), heat release (convective heat transfer, combustion), or differences in concentration (mixing of fluids with different indices of refraction).

A light wave transmitted through the flow with refractive index changes is affected in two different ways: it is deflected from its original direction of propagation and its optical phase is altered in comparison to the phase of the undisturbed wave. In a recording plane at a certain distance behind the flow field under study, three different quantities can be measured (Fig. 5). Each quantity defines a group of optical visualization methods that depend in a different way on the variation of the refractive index, n in the flow field (Table I). A particular method requires the use of an optical apparatus transforming the measurable quantity (light deflection, optical phase changes) into a visual pattern in the recording plane. The pattern is either qualitative (shadowgraph, schlieren) or quantitative (moiré, speckle photography, interferometry), thus allowing for

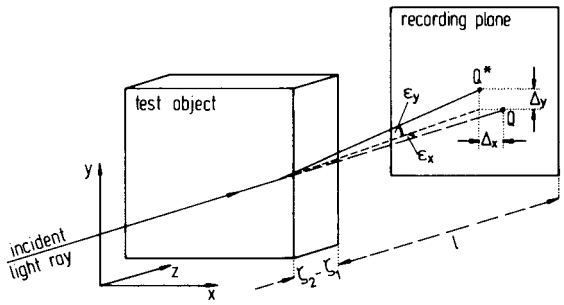


FIGURE 5 Interaction of a light ray with the refractive index field of a fluid flow (test object). The information on light deflection and optical phase changes is recorded in a plane at distance l behind the flow field. [From Merzkirch, W. (1987). "Flow Visualization," 2nd ed., Academic Press, San Diego.]

TABLE I Optical Methods for Line-of-Sight Flow Visualization

Optical method	Quantity measured (see Fig. 5)	Sensitive to changes of	Information on refractive index or density
Shadowgraph	$\overline{\Delta x + \Delta y}$	$\left(\frac{\partial^2 n}{\partial x^2} + \frac{\partial^2 n}{\partial y^2}\right)$	Qualitative
Schlieren	ε_x or ε_y	$\frac{\partial n}{\partial x}$ or $\frac{\partial n}{\partial y}$	Qualitative
Moiré deflectometry	ε_x or ε_y	$\frac{\partial n}{\partial x}$ or $\frac{\partial n}{\partial y}$	Quantitative
Speckle deflectometry	ε_x or ε_y	$\frac{\partial n}{\partial x}$ and $\frac{\partial n}{\partial y}$	Quantitative
Schlieren interferometry	Optical phase change in Q^*	$\frac{\partial n}{\partial x}$ or $\frac{\partial n}{\partial y}$	Quantitative
Reference beam interferometry	Optical phase change in Q^*	n	Quantitative

a deduction of data of the refractive index or density distribution in the flow.

A standard case of application of optical flow visualization is the air flow around a projectile or aerodynamic shape flying at high velocity (Fig. 6). This application to experimental ballistic studies can be traced back to the middle of the 19th century when the methods had been invented in their simplest form. A tremendous push forward in the development of these methods was the availability of laser light, from which particularly the interferometric methods benefitted (Fig. 7). Finally, the combination of holography with interferometry facilitated the mechanical design of practical interferometers.

A major problem with the optical visualization methods, which rely on the transmittance of light (*line-of-sight* methods), is the integration of the information on the refractive index (or density) along the path of the light, in terms of Fig. 5 along the z-direction. The data received in the recording plane are functions of only x and y, whereas in the general case, refractive index or fluid density in the flow are functions of all three space coordinates, x, y, and z.

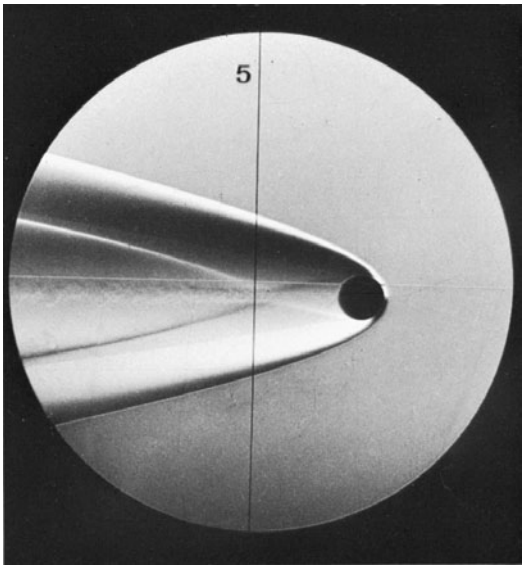


FIGURE 6 Schlieren photograph of a sphere flying at high supersonic velocity from left to right. [From Merzkirch, W. (1987). "Flow Visualization," 2nd ed., Academic Press, San Diego.]

FIGURE 7 Shearing (schlieren) interferogram of the hot gases rising from a candle flame. (Photographed by H. Vanheiden, Universität Essen, Federal Republic of Germany.)

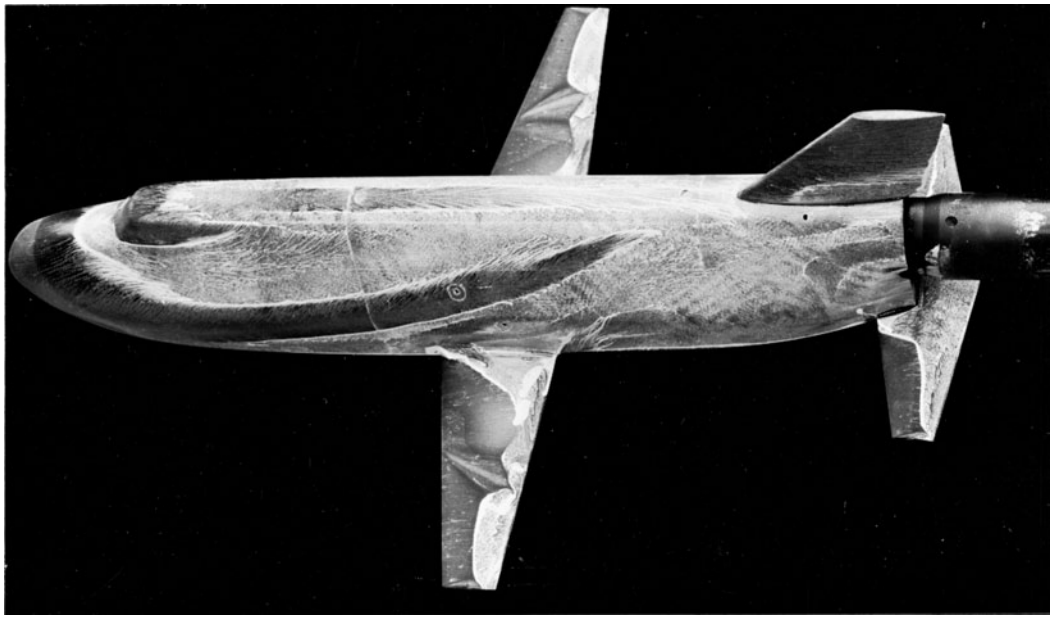


FIGURE 8 Oil film pattern of the flow around an orbiter model that was tested in a wind tunnel. (Courtesy of L. H. Seegmiller, NASA Ames Research Center, Moffet Field, California.)

This three-dimensional distribution of the fluid density can be resolved by recording with the optical setup several projections in different directions through the flow and processing the obtained data with the methods of computer tomography. If the flow has rotational symmetry, one projection only is sufficient, and the axisymmetrically distributed fluid density can be determined by applying to the data an inversion scheme (*Abel inversion*).

IV. SURFACE FLOW VISUALIZATION

The interaction of a fluid flow with the surface of a solid body is a subject of great interest. Many technical measurements are aimed to determine the shear forces, pressure forces, or heating loads applied by the flow to the body. A possible means of estimating the rates of momentum, mass, and heat transfer is to visualize the flow pattern very close to the body surface. For this purpose, the body surface can be coated with a thin layer of a substance that, upon the interaction with the fluid flow, develops a certain visible pattern. This pattern can be interpreted qualitatively, and in some cases, it is possible to measure certain properties of the flow close to the surface. Three different interaction processes can be used for generating different kinds of information.

1. Mechanical Interaction

In the most common technique, which applies to air flows around solid bodies, the solid surface is coated with a thin layer of oil mixed with a finely powdered pigment. Because of frictional forces, the air stream carries the oil with it, and the remaining streaky deposit of the pigment gives information on the direction of the flow close to the surface. The observed pattern may also indicate positions where the flow changes from laminar to turbulent and positions of flow separation and attachment (Fig. 8). Under certain circumstances the wall shear stress can be determined from a measurement of the instantaneous oil film thickness.

2. Chemical Interaction

The solid surface is coated with a substance that changes color upon the chemical reaction with a material with which the flowing fluid is seeded. The reaction, and thereby the color change, is the more intense, the higher the mass transfer from the fluid to the surface. Separated flow regimes with little mass transfer rates can therefore be well discriminated from regions of attached flow. Coating substances are known that change color upon changes of the surface pressure. This is an elegant way for determining pressure loads on the surface of aerodynamic bodies.

3. Thermal Interaction

Coating materials that change color as a function of the surface temperature (temperature sensitive paints, liquid crystals) are known. Observation of the respective color changes allows for determining the instantaneous positions of specific isothermals and deriving the heat

transfer rates to surfaces, which are heated up or cooled down in a fluid flow. Equivalent visible information is available, without the need of surface coating, by applying an infrared camera.

SEE ALSO THE FOLLOWING ARTICLES

FLUID DYNAMICS • HEAT FLOW • IMAGING OPTICS • LIQUIDS, STRUCTURE AND DYNAMICS • IMAGING OPTICS

BIBLIOGRAPHY

Adrian, R. J. (1991). "Particle-imaging techniques for experimental fluid mechanics," *Annu. Rev. Fluid Mechan*. **23,** 261–304.

Emrich, R. J., ed. (1981). "Fluid Dynamics," Vol. 18 of *Methods of Experimental Physics*. Academic Press, New York.

Fomin, N. A. (1998). "Speckle Photography for Fluid Mechanics Measurements," Springer-Verlag, Berlin.

Goldstein, R. J., ed. (1983). "Fluid Mechanics Measurements," Hemisphere, Washington, DC.

Japan Society of Mechanical Engineers, ed. (1988). "Visualized Flow," Pergamon Press, Oxford.

McLachlan, B. G., Kavandi, J. L., Callis, J. B., Gouterman, M., Green, E., Khalil, G., and Burns, D. (1993). "Surface pressure field mapping using luminescent coatings," *Exp. Fluids* **14,** 33–41.

Merzkirch, W. (1987). "Flow Visualization," 2nd ed., Academic Press, San Diego.

Raffel, M., Willert, C., and Kompenhans, J. (1998). "Particle Image Velocimetry," Springer-Verlag, Berlin.

Van Dyke, M., ed. (1982). "An Album of Fluid Motion," Parabolic Press, Stanford, CA.

Yang, W. J., ed. (1989). "Handbook of Flow Visualization," Hemisphere, Washington, DC.

Geodesy

Petr Vaníček
University of New Brunswick

GLOSSARY

Coordinates These are the numbers that define positions in a specific coordinate system. For a coordinate system to be usable (to allow the determination of coordinates) in the real (earth) space, its position and the orientation of its Cartesian axes in the real (earth) space must be known.

Coordinate system In three-dimensional Euclidean space, which we use in geodesy for solving most of the problems, we need either the Cartesian or a curvilinear coordinate system, or both, to be able to work with positions. The Cartesian system is defined by an orthogonal triad of coordinate axes; a curvilinear system is related to its associated generic Cartesian system through some mathematical prescription.

Ellipsoid/spheroid Unless specified otherwise, we understand by this term the geometrical body created by revolving an ellipse around it minor axis, consequently known as an ellipsoid of revolution. By spheroid, we understand a spherelike body, which, of course, includes an ellipsoid as well.

Errors (uncertainties) Inevitable, usually small errors of either a random or systematic nature, which cause uncertainty in every measurement and, consequently, an uncertainty in quantities derived from these observations.

GPS Global Positioning System based on the use of a flock of dedicated satellites.

Gravity anomaly The difference between actual gravity and model gravity, e.g., the normal gravity, where the two gravity values are related to two different points on the same vertical line. These two points have the same value of gravity potential: the actual gravity potential and the model gravity potential, respectively.

Normal gravity field An ellipsoidal model of the real gravity field.

Positioning (static and kinematic) This term is used in geodesy as a synonym for the "determination of positions" of different objects, either stationary or moving.

Satellite techniques Techniques for carrying out different tasks that use satellites and/or satellite-borne instrumentation.

Tides The phenomenon of the earth deformation, including its liquid parts, under the influence of solar and lunar gravitational variations.

WHAT IS GEODESY?

Geodesy is a science, the oldest earth (geo-) science, in fact. It was born of fear and curiosity, driven by a desire to predict natural happenings and calls for the understanding of these happenings. The classical definition, according to one of the "fathers of geodesy" reads: "Geodesy is the science of measuring and portraying the earth's surface" (Helmert, 1880, p. 3). Nowadays, we understand the scope of geodesy to be somewhat wider. It is captured by the following definition (Vaníček and Krakiwsky, 1986, p. 45): "Geodesy is the discipline that deals with the measurement and representation of the earth, including its gravity field, in a three-dimensional time varying space." Note that the contemporary definition includes the study of the earth gravity field (see Section III), as well as studies of temporal changes in positions and in the gravity field (see Section IV).

I. INTRODUCTION

A. Brief History of Geodesy

Little documentation of the geodetic accomplishments of the oldest civilizations, the Sumerian, the Egyptian, the Chinese, and the Indian, has survived. The first firmly documented ideas about geodesy go back to Thales of Miletus (ca. 625–547 BC), Anaximander of Miletus (ca. 611–545 BC), and the school of Pythagoras (ca. 580–500 BC). The Greek students of geodesy included Aristotle (384–322 BC), Eratosthenes (276–194 BC)—the first reasonably accurate determination of the size of the earth, but not taken seriously until 17 centuries later—and Ptolemy (ca. 75–151 AD).

In the Middle Ages, the lack of knowledge of the real size of the earth led Toscanelli (1397–1482) to his famous misinterpretation of the world (Fig. 1), which allegedly lured Columbus to his first voyage west. Soon after, the golden age of exploration got under way and with it the use of position determination by astronomical means. The real extent of the world was revealed to have been close to Eratosthenes's prediction, and people started looking for further quantitative improvements of their conceptual model of the earth. This led to new measurements on the surface of the earth by a Dutchman Snellius (in the 1610s) and a Frenchman Picard (in the 1670s) and the first improvement on Eratosthenes's results. Interested readers can find fascinating details about the oldest geodetic events in Berthon and Robinson (1991).

At about the same time, the notion of the earth's gravity started forming up through the efforts of a Dutchman Stevin (1548–1620), Italians Galileo (1564–1642) and Borelli (1608–1679), an Englishman Horrox (1619–1641), and culminating in Newton's (1642–1727) theory of gravitation. Newton's theory predicted that the earth's globe should be slightly oblate due to the spinning of the earth around its polar axis. A Frenchman Cassini (1625–1712) disputed this prediction; consequently, the French Academy of Science organized two expeditions to Peru and to Lapland under the leadership of Bouguer and

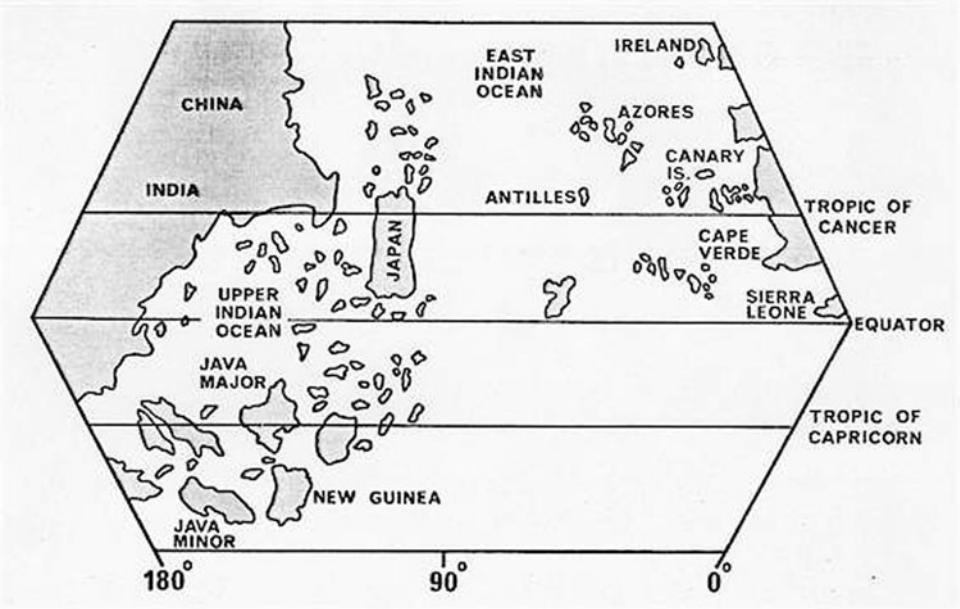


FIGURE 1 Toscanelli's view of the Western Hemisphere.

Maupertuis to measure two meridian arcs. The results confirmed the validity of Newton's prediction. In addition, these measurements gave us the first definition of a meter, as one ten-millionth part of the earth's quadrant.

For 200 years, from about mid-18th century on, geodesy saw an unprecedented growth in its application. Position determination by terrestrial and astronomical means was needed for making maps, and this service, which was naturally provided by geodesists and the image of a geodesist as being only a provider of positions, survives in some quarters till today. In the meantime, the evolution of geodesy as a science continued with contributions by Lagrange (1736–1813), Laplace (1749–1827), Fourier (1768–1830), Gauss (1777–1855), claimed by some geodesists to have been the real founder of geodetic science, Bessel (1784–1846), Coriolis (1792–1843), Stokes (1819–1903), Poincaré (1854–1912), and Albert Einstein. For a description of these contributions, see Vaníček and Krakiwsky (1986, Section 1.3).

B. Geodesy and Other Disciplines and Sciences

We have already mentioned that for more than 200 years geodesy—strictly speaking, only one part of geodesy, i.e., positioning—was applied in mapping in the disguise known on this continent as "control surveying." Positioning finds applications also in the realm of hydrography, boundary demarcation, engineering projects, urban management, environmental management, geography, and planetology. At least one other part of geodesy, geo-kinematic, finds applications also in ecology.

Geodesy has a symbiotic relation with some other sciences. While geodesy supplies geometrical information about the earth, the other geosciences supply physical knowledge needed in geodesy for modeling. Geophysics is the first to come to mind: the collaboration between geophysicists and geodesists is quite wide and covers many facets of both sciences. As a result, the boundary between the two sciences became quite blurred even in the minds of many geoscientists. For example, to some, the study of the global gravity field fits better under geophysics rather than geodesy, while the study of the local gravity field may belong to the branch of geophysics known as exploration geophysics. Other sciences have similar but somewhat weaker relations with geodesy: space science, astronomy (historical ties), oceanography, atmospheric sciences, and geology.

As all exact sciences, geodesy makes heavy use of mathematics, physics, and, of late, computer science. These form the theoretical foundations of geodetic science and, thus, play a somewhat different role vis-à-vis geodesy. In Fig. 2, we have attempted to display the three levels of relations in a cartoon form.

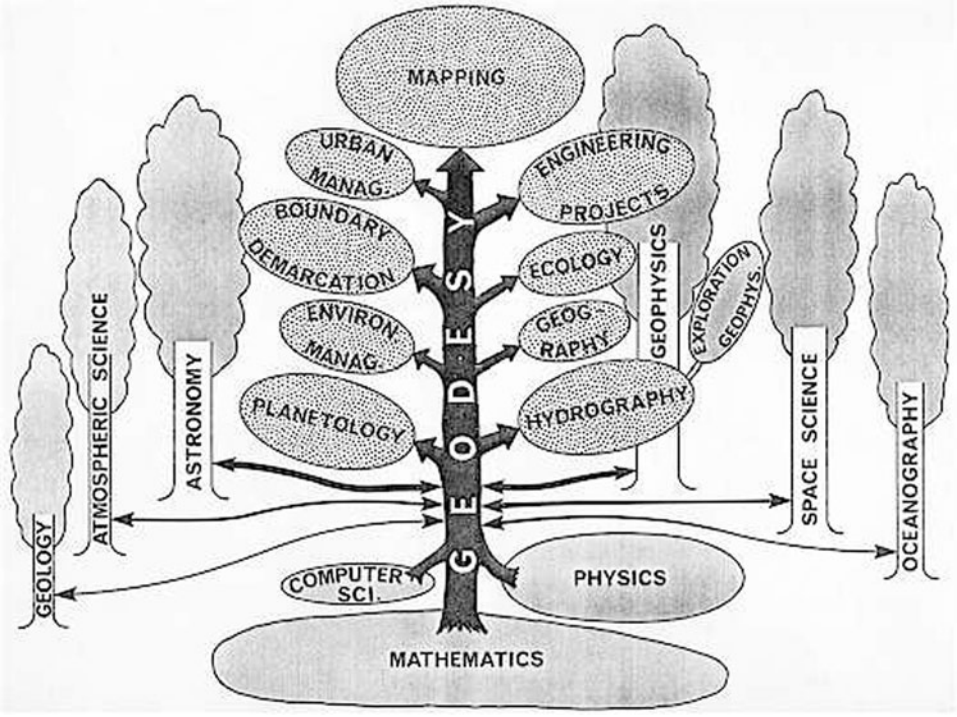


FIGURE 2 Geodesy and other disciplines.

C. Profession and Practice of Geodesy

Geodesy, as most other professions, spans activities ranging from purely theoretical to very applied. The global nature of geodesy dictates that theoretical work be done mostly at universities or government institutions. Few private institutes find it economically feasible to do geodetic research. On the other hand, it is quite usual to combine geodetic theory with practice within one establishment. Much of geodetic research is done under the disguise of space science, geophysics, oceanography, etc.

Of great importance to geodetic theory is international scientific communication. The international organization looking after geodetic needs is the International Association of Geodesy (IAG), the first association of the more encompassing International Union of Geodesy and Geophysics (IUGG) which was set up later in the first third of 20th century. Since its inception, the IAG has been responsible for putting forward numerous important recommendations and proposals to its member countries. It is also operating several international service outfits such as the International Gravimetric Bureau (BGI), the International Earth Rotation Service (IERS), Bureau Internationale des Poids et Mesures—Time Section (BIPM), the International GPS Service (IGS), etc. The interested reader would be well advised to check the current services on the IAG web page.

Geodetic practice is frequently subjugated to mapping needs of individual countries, often only military mapping needs. This results in other components of geodetic work being done under the auspices of other professional institutions. Geodesists practicing positioning are often lumped together with surveyors. They find a limited international forum for exchanging ideas and experience in the International Federation of Surveyors (FIG), a member of the Union of International Engineering Organizations (UIEO).

The educational requirements in geodesy would typically be a graduate degree in geodesy, mathematics, physics, geophysics, etc. for a theoretical geodesist and an undergraduate degree in geodesy, surveying engineering (or geomatics, as it is being called today), survey science, or a similar program for an applied geodesist. Survey technicians, with a surveying (geomatics) diploma from a college or a technological school, would be much in demand for field data collection and routine data manipulations.

II. POSITIONING

A. Coordinate Systems Used in Geodesy

Geodesy is interested in positioning points on the surface of the earth. For this task a well-defined coordinate system is needed. Many coordinate systems are being used in geodesy, some concentric with the earth (geocentric systems), some not. Also, both Cartesian and curvilinear coordinates are used. There are also coordinate systems needed specifically in astronomical and satellite positioning, which are not appropriate to describe positions of terrestrial points in.

Let us discuss the latter coordinate systems first. They are of two distinct varieties: the apparent places and the orbital. The apparent places (AP) and its close relative, the right ascension (RA) coordinate systems, are the ones in which (angular) coordinates of stars are published. The orbital coordinate systems (OR) are designed to be used in describing satellite positions and velocities. The relations between these systems and with the systems introduced below will be discussed in Section II.F. Interested readers can learn about these coordinate systems in Vaníček and Krakiwsky (1986, Chap. 15).

The *geocentric systems* have their z-axis aligned either with the instantaneous spin axis (cf., Section IV.B) of the earth (*instantaneous terrestrial system*) or with a hypothetical spin axis adopted by a convention (*conventional terrestrial systems*). The geocentric systems became useful only quite recently, with the advent of satellite positioning. The nongeocentric systems are used either for local work (observations), in which case their origin would be located at a point on the surface of the earth (topocentric systems called *local astronomic* and *local geodetic*), or for a regional/continental work in the past. These latter nongeocentric (near-geocentric) systems were and are used in lieu of geocentric systems, when these were not yet realizable, and are known as the *geodetic systems*; their origin is usually as close to the center of mass of the earth as the geodesists of yesteryear could make it. They miss the center of mass by anything between a few meters and a few kilometers, and there are some 150 of them in existence around the world.

Both the geocentric and geodetic coordinate systems are used together with *reference ellipsoids* (ellipsoids of revolution or biaxial ellipsoids), also called in some older literature "spheroids." (The modern usage of the term spheroid is for closed, spherelike surfaces, which are more complicated than biaxial ellipsoids.) These reference ellipsoids are taken to be concentric with their coordinate system, geocentric or near geocentric, with the axis of revolution coinciding with the z-axis of the coordinate system. The basic idea behind using the reference ellipsoids is that they fit the real shape of the earth, as described by the geoid (see Section III.B for details) rather well and can thus be regarded as representative, yet simple, expression of the shape of the earth.

The reference ellipsoids are the horizontal surfaces to which the geodetic latitude and longitude are referred,

hence, the name. But to serve in this role, an ellipsoid (together with the associated Cartesian coordinate system) must be fixed with respect to the earth. Such an ellipsoid (fixed with respect to the earth) is often called a *horizontal datum*. In North America we had the North American Datum of 1927, known as NAD 27 (U.S. Department of Commerce, 1973) which was replaced by the geocentric North American Datum of 1983, referred to as NAD 83 (Boal and Henderson, 1988; Schwarz, 1989).

The horizontal geodetic coordinates, *latitude* φ and *longitude* λ, together with the *geodetic height* h (called by some authors by ellipsoidal height, a logical *nonsequitur*, as we shall see later), make the basic triplet of curvilinear coordinates widely used in geodesy. They are related to their associated Cartesian coordinates x, y, and z by the following simple expressions:

$$\begin{aligned} x &= (N+h)\cos\varphi\cos\lambda \\ y &= (N+h)\cos\varphi\sin\lambda \\ z &= (Nb^2/a^2+h)\sin\varphi, \end{aligned} \tag{1}$$

where N is the local radius of curvature of the reference ellipsoid in the east–west direction,

$$N = a^2\,(a^2\cos^2\varphi + b^2\sin^2\varphi)^{-1/2}, \tag{2}$$

where a is the major semi-axis and b is the minor semi-axis of the reference ellipsoid. We note that the geodetic heights are not used in practice; for practical heights, see Section II.B. It should be noted that the horizontal geodetic coordinates are the ones that make the basis for all maps, charts, legal land and marine boundaries, marine and land navigation, etc. The transformations between these horizontal coordinates and the two-dimensional Cartesian coordinates x, y on the maps are called *cartographic mappings*.

Terrestrial (geocentric) coordinate systems are used in satellite positioning. While the instantaneous terrestrial (IT) system is well suited to describe instantaneous positions in, the conventional terrestrial (CT) systems are useful for describing positions for archiving. The conventional terrestrial system recommended by IAG is the International Terrestrial Reference System (ITRS), which is "fixed to the earth" via several permanent stations whose horizontal tectonic velocities are monitored and recorded.

gets associated with the time of fixing by time tagging. The "realization" of the ITRS by means of coordinates of some selected points is called the International Terrestrial Reference Frame (ITRF). Transformation parameters needed for transforming coordinates from one epoch to the next are produced by the International Earth Rotation Service (IERS) in Paris, so one can keep track of the time evolution of the positions. For more detail the reader is referred to the web site of the IERS or to a popular article by Boucher and Altamini (1996).

B. Point Positioning

It is not possible to determine either three-dimensional (3D) or two-dimensional (2D) (horizontal) positions of isolated points on the earth's surface by terrestrial means. For point positioning we must be looking at celestial objects, meaning that we must be using either optical techniques to observe stars [geodetic astronomy, see Mueller (1969)] or electronic/optical techniques to observe earth's artificial satellites (satellite positioning, cf., Section V.B). Geodetic astronomy is now considered more or less obsolete, because the astronomically determined positions are not very accurate (due to large effects of unpredictable atmospheric refraction) and also because they are strongly affected by the earth's gravity field (cf., Section III.D). Satellite positioning has proved to be much more practical and more accurate.

On the other hand, it is possible to determine heights of some isolated points through terrestrial means by tying these points to the sea level. Practical heights in geodesy, known as *orthometric heights* and denoted by H^{O}, or simply by H, are referred to the geoid, which is an equipotential surface of the earth's gravity field (for details, see Section III.B) approximated by the mean sea level (MSL) surface to an accuracy of within ± 1.5 m. The difference between the two surfaces arises from the fact that seawater is not homogeneous and because of a variety of dynamical effects on the seawater. The height of the MSL above the geoid is called the *sea surface topography* (SST). It is a very difficult quantity to obtain from any measurements; consequently, it is not yet known very accurately. We note that the orthometric height H is indeed different from the geodetic height h discussed in Section II.A: the relation between the two kinds of heights is shown in Fig. 3, where the quantity N, the height of the geoid above the reference ellipsoid, is usually called the *geoidal height* (geoid

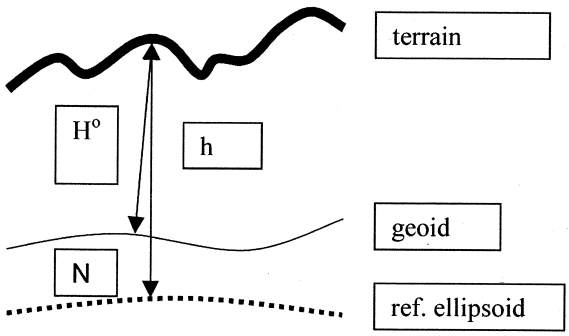


FIGURE 3 Orthometric and geodetic heights.

undulation) (cf., Section III.B). Thus, the knowledge of the geoid is necessary for transforming the geodetic to orthometric heights and vice versa. We note that the acceptance of the standard geodetic term of "geoidal height" (height of the geoid above the reference ellipsoid) makes the expression "ellipsoidal height" for (geodetic) height of anything above the reference ellipsoid, a logical *non-sequitur* as pointed out above.

We have seen above that the geodetic height is a purely geometrical quantity, the length of the normal to the reference ellipsoid between the ellipsoid and the point of interest. The orthometric height, on the other hand, is defined as the length of the plumbline (a line that is always normal to the equipotential surface of the gravity field) between the geoid and the point of interest and, as such, is intimately related to the gravity field of the earth. (As the plumbline is only slightly curved, the length of the plumbline is practically the same as the length of the normal to the geoid between the geoid and the point of interest. Hence, the equation $h \cong H + N$ is valid everywhere to better than a few millimeters.) The defining equation for the orthometric height of point A (given by its position vector $\mathbf{r}_A$) is

$$H^O(\mathbf{r}_A) = H(\mathbf{r}_A) = [W_0 - W(\mathbf{r}_A)]/\text{mean}(g_A), \quad (3)$$

where W_0 stands for the constant gravity potential on the geoid, $W(\mathbf{r}_A)$ is the gravity potential at point A, and mean(g_A) is the mean value of gravity (for detailed treatment of these quantities, see Sections III.A and III.B) between A and the geoid—these. From these equations it can be easily gleaned that orthometric heights are indeed referred to the geoid (defined as $W_0 = 0$). The mean(g) cannot be measured and has to be estimated from gravity observed at A, $g(\mathbf{r}_A)$, assuming a reasonable value for the vertical gradient of gravity within the earth. Helmert (1880) hypothesized the value of 0.0848 mGal m^{-1} suggested independently by Poincaré and Prey to be valid for the region between the geoid and the earth's surface (see Section III.C), to write

$$\text{mean}(g_A) \cong g(\mathbf{r}_A) + 0.0848\, H(\mathbf{r}_A)/2\,[\text{mGal}]. \quad (4)$$

For the definition of units of gravity, Gal, see Section III.A. Helmert's (approximate) orthometric heights are used for mapping and for technical work almost everywhere. They may be in error by up to a few decimeters in the mountains. Equipotential surfaces at different heights are not parallel to the equipotential surface at height 0, i.e., the geoid. Thus, orthometric heights of points on the same equipotential surface $W = \text{const.} \neq W_0$ are generally not the same and, for example, the level of a lake appears to be sloping. To avoid this, and to allow the physical laws to be taken into proper account, another system of height is used: *dynamic heights*. The dynamic height of point A is defined as

$$H^D(\mathbf{r}_A) = [W_0 - W(\mathbf{r}_A)]/\gamma_{\text{ref}}, \quad (5)$$

where γ_{ref} is a selected (reference) value of gravity, constant for the area of interest. We note that points on the same equipotential surface have the same dynamic height; that dynamic heights are referred to the geoid but they must be regarded as having a scale that changes from point to point.

We must also mention the third most widely used height system, the *normal heights*. These heights are defined by

$$H^N(\mathbf{r}_A) = H^*(\mathbf{r}_A) = [W_0 - W(\mathbf{r}_A)]/\text{mean}(\gamma_A), \quad (6)$$

where mean(γ_A) is the value of the model gravity called "normal" (for a detailed explanation, see Section III.A) at a height of $H^N(\mathbf{r}_A)/2$ above the reference ellipsoid along the normal to the ellipsoid (Molodenskij, Eremeev, and Yurkina, 1960). We refer to this value as mean because it is evaluated at a point halfway between the reference ellipsoid and the locus of $H^N(\mathbf{r}_A)$, referred to the reference ellipsoid, which (locus) surface is called the *telluroid*. For practical purposes, normal heights of terrain points A are referred to a different surface, called *quasi-geoid* (cf., Section III.G), which, according to Molodenskij, can be computed from gravity measurements in a similar way to the computation of the geoid.

C. Relative Positioning

Relative positioning, meaning positioning of a point with respect to an existing point or points, is the preferred mode of positioning in geodesy. If there is intervisibility between the points, terrestrial techniques can be used. For satellite relative positioning, the intervisibility is not a requirement, as long as the selected satellites are visible from the two points in question. The accuracy of such relative positions is usually significantly higher than the accuracy of single point positions.

The classical terrestrial techniques for 2D relative positioning make use of angular (horizontal) and distance measurements, which always involve two or three points. These techniques are thus differential in nature. The computations of the relative 2D positions are carried out either on the horizontal datum (reference ellipsoid), in terms of latitude difference $\Delta\varphi$ and longitude difference $\Delta\lambda$, or on a map, in terms of Cartesian map coordinate differences Δx and Δy. In either case, the observed angles, azimuths, and distances have to be first transformed (reduced) from the earth's surface, where they are acquired, to the reference ellipsoid, where they are either used in the computations or transformed further onto the selected mapping plane. We shall not explain these reductions here; rather, we would advise the interested reader to consult one

of the classical geodetic textbooks (e.g., Zakatov, 1953; Bomford, 1971).

To determine the relative position of one point with respect to another on the reference ellipsoid is not a simple proposition, since the computations have to be carried out on a curved surface and Euclidean geometry no longer applies. Links between points can no longer be straight lines in the Euclidean sense; they have to be defined as geodesics (the shortest possible lines) on the reference ellipsoid. Consequently, closed form mathematical expressions for the computations do not exist, and use has to be made of various series approximations. Many such approximations had been worked out, which are valid for short, medium, or long geodesics. For 200 years, coordinate computations on the ellipsoid were considered to be the backbone of (classical) geodesy, a litmus test for aspiring geodesists. Once again, we shall have to desist from explaining the involved concepts here as there is no room for them in this small article. Interested readers are referred once more to the textbooks cited above.

Sometimes, preference is given to carrying out the relative position computations on the mapping plane, rather than on the reference ellipsoid. To this end, a suitable cartographic mapping is first selected, normally this would be the conformal mapping used for the national/state geodetic work. This selection carries with it the appropriate mathematical mapping formulae and distortions associated with the selected mapping (Lee, 1976). The observed angles ω, azimuths α, and distances S (that had been first reduced to the reference ellipsoid) are then reduced further (distorted) onto the selected mapping plane where (2D) Euclidean geometry can be applied. This is shown schematically in Fig. 4. Once these reductions have been carried out, the computation of the (relative) position of the unknown point B with respect to point A already known on the mapping plane is then rather trivial:

$$x_B = x_A + \Delta x_{AB}, \qquad y_B = y_A + \Delta y_{AB}. \tag{7}$$

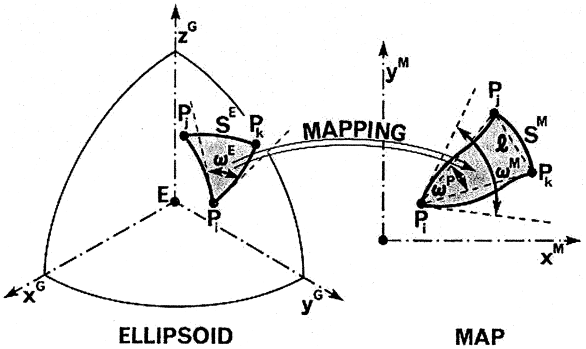


FIGURE 4 Mapping of ellipsoid onto a mapping plane.

Relative vertical positioning is based on somewhat more transparent concepts. The process used for determining the height difference between two points is called *geodetic levelling* (Bomford, 1971). Once the levelled height difference is obtained from field observations, one has to add to it a small correction based on gravity values along the way to convert it to either the orthometric, the dynamic, or the normal height difference. Geodetic levelling is probably the most accurate geodetic relative positioning technique. To determine the geodetic height difference between two points, all we have to do is to measure the vertical angle and the distance between the points. Some care has to be taken that the vertical angle is reckoned from a plane perpendicular to the ellipsoidal normal at the point of measurement.

Modern extraterrestrial (satellite and radio astronomical) techniques are inherently three dimensional. Simultaneous observations at two points yield 3D coordinate differences that can be added directly to the coordinates of the known point A on the earth's surface to get the sought coordinates of the unknown point B (on the earth's surface). Denoting the triplet of Cartesian coordinates (x, y, z) in any coordinate system by $\mathbf{r}$ and the triplet of coordinate differences $(\Delta x, \Delta y, \Delta z)$ by $\Delta\mathbf{r}$, the 3D position of point B is given simply by

$$\mathbf{r}_B = \mathbf{r}_B + \Delta\mathbf{r}_{AB}, \tag{8}$$

where $\Delta\mathbf{r}_{AB}$ comes from the observations.

We shall discuss in Section V.B how the "base vector" $\Delta\mathbf{r}_{AB}$ is derived from satellite observations. Let us just mention here that $\Delta\mathbf{r}_{AB}$ can be obtained also by other techniques, such as radio astronomy, inertial positioning, or simply from terrestrial observations of horizontal and vertical angles and distances. Let us show here the principle of the interesting radio astronomic technique for the determination of the base vector, known in geodesy as *Very Long Baseline Interferometry* (VLBI). Figure 5 shows schematically the pair of radio telescopes (steerable antennas, A and B) following the same quasar whose celestial position is known (meaning that $\mathbf{e}_s$ is known). The time delay τ can be measured very accurately and the base vector $\Delta\mathbf{r}_{AB}$ can be evaluated from the following equation:

$$\tau = c^{-1}\,\mathbf{e}_s \Delta\mathbf{r}_{AB}, \tag{9}$$

where c is the speed of light. At least three such equations are needed for three different quasars to solve for $\Delta\mathbf{r}_{AB}$.

Normally, thousands of such equations are available from dedicated observational campaigns. The most important contribution of VLBI to geodesy (and astronomy) is that it works with directions (to quasars) which can be considered as the best approximations of directions in an inertial space.

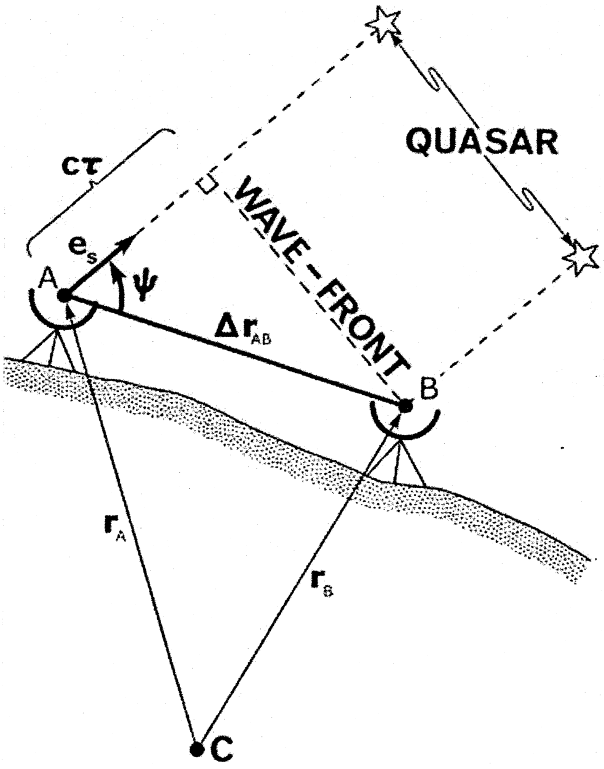


FIGURE 5 Radioastronomical interferometry.

D. Geodetic Networks

In geodesy we prefer to position several points simultaneously because when doing so we can collect redundant information that can be used to check the correctness of the whole positioning process. Also, from the redundancy, one can infer the internal consistency of the positioning process and estimate the accuracy of so determined positions (cf., Section II.E). Thus, the classical geodetic way of positioning points has been in the mode of *geodetic networks*, where a whole set of points is treated simultaneously. This approach is, of course, particularly suitable for the terrestrial techniques that are differential in nature, but the basic rationale is equally valid even for modern positioning techniques. After the observations have been made in the field, the positions of network points are estimated using optimal estimation techniques that minimize the quadratic norm of observation residuals from systems of (sometimes hundreds of thousands) overdetermined (observation) equations. A whole body of mathematical and statistical techniques dealing with network design and position estimation (*network adjustment*) has been developed; the interested reader may consult Grafarend and Sansò (1985), Hirvonen (1971), and Mikhail (1976) for details.

The 2D (horizontal) and 1D (vertical) geodetic networks of national extent, sometimes called national control networks, have been the main tool for positioning needed in mapping, boundary demarcation, and other geodetic applications. For illustration, the Canadian national geodetic levelling network is shown in Fig. 6. We note that national networks are usually interconnected to create continental networks that are sometimes adjusted together—as is the case in North America—to make the networks more homogeneous. Local networks in one, two, and three dimensions have been used for construction purposes. In classical geodetic practice, the most widely encountered networks are horizontal, while 3D networks are very rare.

Vertical (height, levelling) networks are probably the best example of how differential positioning is used together with the knowledge of point heights in carrying the height information from the seashore inland. The heights of selected shore benchmarks are first derived from the observations of the sea level (cf., Section II.B), carried out by means of *tide gauges* (also known in older literature as mareographs) by means of short levelling lines. These basic benchmarks are then linked to the network by longer levelling lines that connect together a whole multitude of land benchmarks (cf., Fig. 6).

Modern satellite networks are inherently three-dimensional. Because the intervisibility is not a requirement for relative satellite positioning, satellite networks can and do contain much longer links and can be much larger in geographical extent. Nowadays, global geodetic networks are constructed and used for different applications.

E. Treatment of Errors in Positions

All positions, determined in whatever way, have errors, both systematic and random. This is due to the fact that every observation is subject to an error; some of these errors are smaller, some are larger. Also, the mathematical models from which the positions are computed are not always completely known or properly described. Thus, when we speak about positions in geodesy, we always mention the accuracy/error that accompanies it. How are these errors expressed?

Random errors are described by following quadratic form:

$$\xi^{\mathrm{T}}\mathbf{C}^{-1}\xi = C_\alpha, \tag{10}$$

where $\mathbf{C}$ is the *covariance matrix* of the position (a three by three matrix composed of variances and covariances of the three coordinates which comes as a by-product of the network adjustment) and C_α is a factor that depends on the probability density function involved in the position estimation and on the desired probability level α. This

FIGURE 6 The Canadian national geodetic levelling network (Source: www.nrcan.gc.ca. Copyright: Her Majesty the Queen in Right of Canada, Natural Resources Canada, Geodetic Survey Division. All rights reserved.)

quadratic form can be interpreted as an equation of an ellipsoid, called a confidence region in statistics or an *error ellipsoid* in geodetic practice. The understanding is that if we know the covariance matrix $\mathbf{C}$ and select a probability level α we are comfortable with, then the vector difference ξ between the estimated position $\mathbf{r}^*$ and the real position $\mathbf{r}$ is, with probability α, within the confines of the error ellipsoid.

The interpretation of the error ellipsoid is a bit tricky. The error ellipsoid described above is called *absolute*, and one may expect that errors (and thus also accuracy) thus measured refer to the coordinate system in which the positions are determined. They do not! They actually refer to the point (points) given to the network adjustment (cf., Section II.D) for fixing the position of the adjusted point configuration. This point (points) is sometimes called the "datum" for the adjustment, and we can say that the absolute confidence regions are really relative with respect to the "adjustment datum." As such, they have a natural tendency to grow in size with the growing distance of the point of interest from the adjustment datum. This behavior curtails somewhat the usefulness of these measures.

Hence, in some applications, *relative* error ellipsoids (confidence regions) are sought. These measure errors (accuracy) of one position, A, with respect to another position, B, and thus refer always to pairs of points. A relative confidence region is defined by an expression identical to Eq. (10), except that the covariance matrix used, $\mathbf{C}_{\Delta AB}$, is that of the three coordinate differences $\Delta\mathbf{r}_{AB}$ rather than the three coordinates $\mathbf{r}$. This covariance matrix is evaluated from the following expression:

$$\mathbf{C}_{\Delta AB} = \mathbf{C}_A + \mathbf{C}_B - \mathbf{C}_{AB} - \mathbf{C}_{BA}, \qquad (11)$$

where $\mathbf{C}_A$ and $\mathbf{C}_B$ are the covariance matrices of the two points, A and B, and $\mathbf{C}_{AB} = \mathbf{C}_{BA}^T$ is the cross-covariance matrix of the two points. The cross-covariance matrix comes also as a by-product of the network adjustment. It should be noted that the cross-covariances (cross-correlations) between the two points play a very significant role here: when the cross-correlations are strong, the relative confidence region is small and vice versa.

When we deal with 2D instead of 3D coordinates, the confidence regions (absolute and relative) also become two dimensional. Instead of having error ellipsoids, we have error ellipses—see Fig. 7 that shows both absolute and relative error ellipses as well as errors in the distance S, σ_S and azimuth α, $S\sigma_\alpha$, computed (estimated) from the positions of A and B. In the 1D case (heighting), confidence regions degenerate to line segments. The Dilution of Precision (DOP) indicators used in GPS (cf., Section V.B)

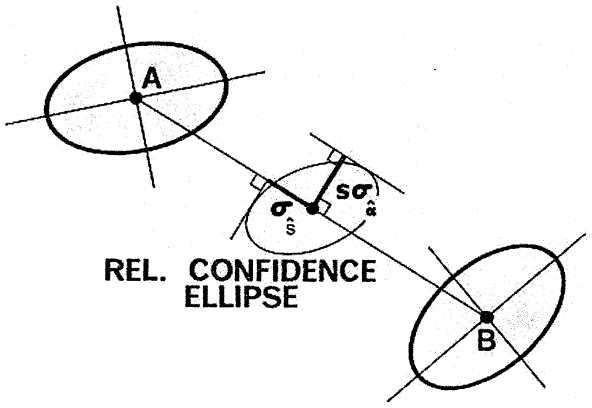


FIGURE 7 Absolute and relative error ellipses.

are related to the idea of (somewhat simplified) confidence regions.

Once we know the desired confidence region(s) in one coordinate system, we can derive the confidence region in any other coordinate system. The transformation works on the covariance matrix used in the defining expression (10) and is given by

$$\mathbf{C}^{(2)} = \mathbf{T}\mathbf{C}^{(1)}\mathbf{T}^{\mathrm{T}}, \tag{12}$$

where $\mathbf{T}$ is the Jacobian of transformation from the first to the second coordinate systems evaluated for the point of interest, i.e., $\mathbf{T} = \mathbf{T}(\mathbf{r})$.

Systematic errors are much more difficult to deal with. Evaluating them requires an intimate knowledge of their sources, and these are not always known. The preferred way of dealing with systematic errors is to prevent them from occurring in the first place. If they do occur, then an attempt is made to eliminate them as much as possible.

There are other important issues that we should discuss here in connection with position errors. These include concepts of blunder elimination, reliability, geometrical strength of point configurations, and more. Unfortunately, there is no room to get into these concepts here, and the interested reader may wish to consult Vaníček and Krakiwsky (1986) or some other geodetic textbook.

F. Coordinate Transformations

A distinction should be made between (abstract) "coordinate system transformations" and "coordinate transformations": coordinate systems do not have any errors associated with them while coordinates do. The transformation between two Cartesian coordinate systems [first (1) and second (2)] can be written in terms of hypothetical positions $\mathbf{r}^{(1)}$ and $\mathbf{r}^{(2)}$ as

$$\mathbf{r}^{(2)} = \mathbf{R}(\varepsilon_x, \varepsilon_y, \varepsilon_z,)\mathbf{r}^{(1)} + \mathbf{t}_0^{(2)}, \tag{13}$$

where $\mathbf{R}(\varepsilon_x, \varepsilon_y, \varepsilon_z,)$ is the "rotation matrix," which after application to a vector rotates the vector by the three *misalignment angles* $\varepsilon_x, \varepsilon_y, \varepsilon_z$, around the coordinate axes, and $\mathbf{t}_0^{(2)}$ is the position vector of the origin of the first system reckoned in the second system, called the *translation vector*.

The transformation between coordinates must take into account the errors in both coordinates/coordinate sets (in the first and second coordinate system), particularly the systematic errors. A transformation of coordinates thus consists of two distinct components: the transformation between the corresponding coordinate systems as described above, plus a model for the difference between the errors in the two coordinate sets. The standard illustration of such a model is the inclusion of the scale factor, which accounts for the difference in linear scales of the two coordinate sets. In practice, when dealing with coordinate sets from more extensive areas such as states or countries, these models are much more elaborate, as they have to model the differences in the deformations caused by errors in the two configurations. These models differ from country to country. For unknown reasons, some people prefer not to distinguish between the two kinds of transformations.

Figure 8 shows a commutative diagram for transformations between most of the coordinate systems used in geodesy. The quantities in rectangles are the transformation parameters, the misalignment angles, and translation components. For a full understanding of the involved transformations, the reader is advised to consult Vaníček and Krakiwsky (1986, Chap. 15).

Let us just mention that sometimes we are not interested in transforming positions (position vectors, triplets of coordinates), but small (differential) changes $\delta\mathbf{r}$ in positions $\mathbf{r}$ as we saw in Eq. (12). In this case, we are not concerned with translations between the coordinate systems, only misalignments are of interest. Instead of using the rotation matrix, we may use the Jacobian of transformation, getting

$$\delta\mathbf{r}^{(2)}(\mathbf{r}) = \mathbf{T}(\mathbf{r})\,\delta\mathbf{r}^{(1)}(\mathbf{r}). \tag{14}$$

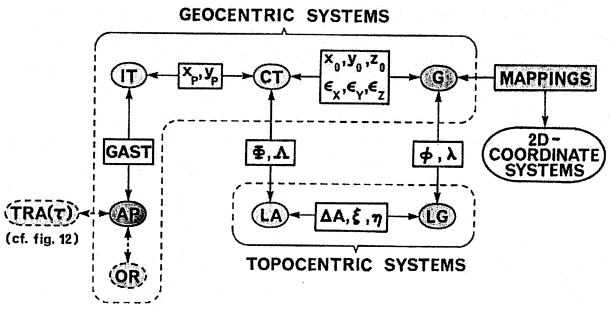


FIGURE 8 Commutative diagram of transformations between coordinate systems.

The final topic we want to discuss in this section is one that we are often faced with in practice: given two corresponding sets of positions (of the same points) in two different coordinate systems we wish to determine the transformation parameters for the two systems. This is done by means of Eq. (13), where $\varepsilon_x, \varepsilon_y, \varepsilon_z, \mathbf{t}_1^{(2)}$ become the six unknown transformation parameters, while $\mathbf{r}^{(1)}$, $\mathbf{r}^{(2)}$ are the known quantities. The set of known positions has to consist of at least three noncollinear points so we get at least six equations for determining the six unknowns. We must stress that the set of position vectors $\mathbf{r}^{(1)}$, $\mathbf{r}^{(2)}$ has to be first corrected for the distortions caused by the errors in both sets of coordinates. These distortions are added back on if we become interested in coordinate transformation.

G. Kinematic Positioning and Navigation

As we have seen so far, classical geodetic positioning deals with stationary points (objects). In recent times, however, geodetic positioning has found its role also in positioning moving objects such as ships, aircraft, and cars. This application became known as *kinematic positioning*, and it is understood as being the real-time positioning part of navigation. Formally, the task of kinematic positioning can be expressed as

$$\mathbf{r}(t) = \mathbf{r}(t_0) + \int_{t_0}^{t} \mathbf{v}(t)\, dt, \tag{15}$$

where t stands for time and $\mathbf{v}(t)$ is the observed change in position in time, i.e., velocity (vector) of the moving object. The velocity vector can be measured on the moving vehicle in relation to the surrounding space or in relation to an inertial coordinate system by an inertial positioning system. We note that, in many applications, the attitude (roll and pitch) of the vehicle is also of interest.

Alternatively, optical astronomy or point satellite positioning produce directly the string of positions, $\mathbf{r}(t_1)$, $\mathbf{r}(t_2), \ldots, \mathbf{r}(t_n)$, that describe the required *trajectory* of the vehicle, without the necessity of integrating over velocities. Similarly, a relative positioning technique, such as the hyperbolic radio system Loran-C (or Hi-Fix, Decca, Omega in the past), produces a string of position changes, $\Delta\mathbf{r}(t_0, t_1)$, $\Delta\mathbf{r}(t_1, t_2), \ldots, \Delta\mathbf{r}(t_{n-1}, t_n)$, which once again define the trajectory. We note that these techniques are called hyperbolic because positions or position differences are determined from intersections of hyperbolae, which, in turn, are the loci of constant distance differences from the land-located radiotransmitters. Relative satellite positioning is also being used for kinematic positioning, as we shall see later in Section V.B.

For a navigator, it is not enough to know his position $\mathbf{r}(t_n)$ at the time t_n. The navigator must also have the position estimates for the future, i.e., for the times $t_n, t_{n+1}, \ldots,$ to be able to navigate safely, he has to have the *predicted positions*. Thus, the kinematic positioning described above has to be combined with a *navigation algorithm*, a predictive filter which predicts positions in the future based on the observed position in the past, at times $t_1, t_2, \ldots, t_n$. Particularly popular seem to be different kinds of Kalman's filters, which contain a feature allowing one to describe the dynamic characteristics of the vehicle navigating in a specified environment. Kalman's filters do have a problem with environments that behave in an unpredictable way, such as an agitated sea. We note that some of the navigation algorithms accept input from two or more kinematic position systems and combine the information in an optimal way.

In some applications, it is desirable to have a post-mission record of trajectories available for future retracing of these trajectories. These post-mission trajectories can be made more accurate than the real-time trajectories (which, in turn, are of course more accurate than the predicted trajectories). Most navigation algorithms have the facility of *post-mission smoothing* the real-time trajectories by using all the data collected during the mission.

III. EARTH'S GRAVITY FIELD

A. Origin of the Earth's Gravity Field

In geodesy, we are interested in studying the gravity field in the macroscopic sense where the quantum behavior of gravity does not have to be taken into account. Also, in terrestrial gravity work, we deal with velocities that are very much smaller than the speed of light. Thus, we can safely use Newtonian physics and may begin by recalling mass attraction force $\mathbf{f}$ defined by Newton's integral

$$\mathbf{f}(\mathbf{r}) = \mathbf{a}(\mathbf{r})m = \left(G \int_{\boldsymbol{B}} \rho(\mathbf{r}')(\mathbf{r}' - \mathbf{r})|\mathbf{r}' - \mathbf{r}|^{-3}\, dV \right) m, \tag{16}$$

where $\mathbf{r}$ and $\mathbf{r}'$ are position vectors of the point of interest and the dummy point of the integration; $\boldsymbol{B}$ is the attracting massive body of density ρ, i.e., the earth; V stands for volume; G is Newton's gravitational constant; m is the mass of the particle located at $\mathbf{r}$; and $\mathbf{a}(\mathbf{r})$ is the acceleration associated with the particle located at $\mathbf{r}$ (see Fig. 9). We can speak about the acceleration $\mathbf{a}(\mathbf{r})$, called *gravitation*, even when there is no mass particle present at $\mathbf{r}$, but we cannot measure it (only an acceleration of a mass can be measured). This is the idea behind the definition of the *gravitational field* of body $\boldsymbol{B}$, the earth; this field is defined at all points $\mathbf{r}$. The physical units of gravitation are those of an acceleration, i.e., m s^{-2}; in practice, units of cm s^{-2},

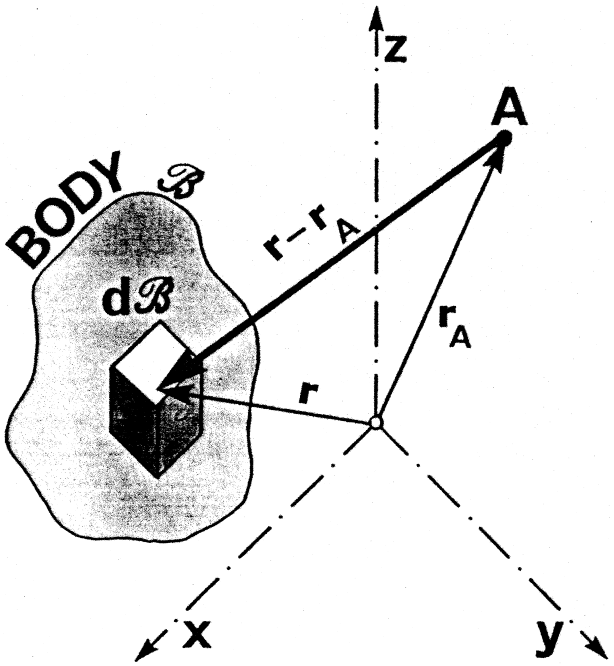


FIGURE 9 Mass attraction.

called "Gal" [to commemorate Galileo's (c.f., Section I.A) contribution to geodesy], are often used.

Newton's gravitational constant G represents the ratio between mass acting in the "attracted capacity" and the same mass acting in the "attracting capacity." From Eq. (16) we can deduce the physical units of G, which are 10 $\mathrm{kg}^{-1}\ \mathrm{m}^3\ \mathrm{s}^{-2}$. The value of G has to be determined experimentally. The most accurate measurements are obtained from tracking deep space probes that move in the gravitational field of the earth. If a deep space probe is sufficiently far from the earth (and the attractions of the other celestial bodies are eliminated mathematically), then the physical dimensions of the probe become negligible. At the same time, the earth can be regarded with sufficient accuracy as a sphere with a laterally homogeneous density distribution. Under these circumstances, the gravitational field of the earth becomes radial, i.e., it will look as if it were generated by a particle of mass M equal to the total mass of the earth:

$$M = \int_B \rho(\mathbf{r}')\, dV. \tag{17}$$

When a "geocentric" coordinate system is used in the computations, the probe's acceleration becomes

$$\mathbf{a}(\mathbf{r}) = -GM\mathbf{r}|\mathbf{r}|^{-3}. \tag{18}$$

Thus, the gravitational constant G, or more accurately GM, called the *geocentric constant*, can be obtained from purely geometrical measurements of the deep space probe positions $\mathbf{r}(t)$. These positions, in turn, are determined from measurements of the propagation of electromagnetic waves and as such depend very intimately on the accepted value of the speed of light c. The value of GM is now thought to be $(3{,}986{,}004.418 \pm 0.008) * 10^8\ \mathrm{m}^3\ \mathrm{s}^{-2}$ (Ries *et al.*, 1992), which must be regarded as directly dependent on the accepted value of c. Dividing the geocentric constant by the mass of the earth $[(5.974 \pm 0.001) * 10^{24}\ \mathrm{kg}]$, one obtains the value for G as $(6.672 \pm 0.001) * 10^{-11}\ \mathrm{kg}^{-1}\ \mathrm{m}^3\ \mathrm{s}^{-2}$.

The earth spins around its instantaneous spin axis at a more or less constant angular velocity of once per "sidereal day"—sidereal time scale is taken with respect to fixed stars, which is different from the solar time scale, taken with respect to the sun. This spin gives rise to a centrifugal force that acts on each and every particle within or bound with the earth. A particle, or a body, which is not bound with the earth, such as an earth satellite, is not subject to the centrifugal force. This force is given by the following equation:

$$\mathbf{F}(\mathbf{r}) = \omega^2\, \mathbf{p}(\mathbf{r})m, \tag{19}$$

where $\mathbf{p}(\mathbf{r})$ is the projection of $\mathbf{r}$ onto the equatorial plane, ω is the siderial angular velocity of 1 revolution per day $(7.292115 * 10^{-5}\ \mathrm{rad}\ \mathrm{s}^{-1})$, and m is the mass of the particle subjected to the force. Note that the particles on the spin axis of the earth experience no centrifugal force as the projection $\mathbf{p}(\mathbf{r})$ of their radius vector equals to $\mathbf{0}$. We introduce the *centrifugal acceleration* $\mathbf{a}_c(\mathbf{r})$ at point $\mathbf{r}$ as $\omega^2\, \mathbf{p}(\mathbf{r})$ and speak of the centrifugal acceleration field much in the same way we speak of the gravitational field (acceleration) $\mathbf{a}(\mathbf{r})$.

The earth (B) gravitation is denoted by $\mathbf{g}_g$ (subscripted g for gravitation) rather than $\mathbf{a}$, and its centrifugal acceleration is denoted by $\mathbf{g}_c$ (c for centrifugal) rather than $\mathbf{a}_c$. When studying the fields $\mathbf{g}_g$ and $\mathbf{g}_c$ acting at points bound with the earth (spinning with the earth), we normally lump these two fields together and speak of the earth's *gravity field* $\mathbf{g}$:

$$\mathbf{g}(\mathbf{r}) = \mathbf{g}_g(\mathbf{r}) + \mathbf{g}_c(\mathbf{r}). \tag{20}$$

A stationary test mass m located at any of these points will sense the total gravity vector $\mathbf{g}$ (acceleration).

If the (test) mass moves with respect to the earth, then another (virtual) force affects the mass: the Coriolis force, responsible, for instance, for the geostrophic motion encountered in air or water movement. In the studies of the earth's gravity field, Coriolis' force is not considered. Similarly, temporal variation of the gravity field, due to variations in density distribution and in earth's rotation speed, which are small compared to the magnitude of the field itself, is mostly not considered either. It is studied separately within the field of geo-kinematics (Section IV).

B. Gravity Potential

When we move a mass m in the gravity field $\mathbf{g}(\mathbf{r})$ from location $\mathbf{r}_1$ to location $\mathbf{r}_2$, to overcome the force $\mathbf{g}(\mathbf{r})\,m$ of the field, we have to do some work w. This work is expressed by the following line integral:

$$w = -\int_{r_1}^{r_2} \mathbf{g}(\mathbf{r})' m \, d\mathbf{r}'. \tag{21}$$

Note that the physical units of work w are kg m^2 s^{-2}. Fortunately, for the gravitational field the amount of work does not depend on what trajectory is followed when moving the particle from $\mathbf{r}_1$ to $\mathbf{r}_2$. This property can be expressed as

$$\oint_C \mathbf{g}(\mathbf{r}')\, d\mathbf{r}' m = \oint_C \mathbf{g}(\mathbf{r}')\, d\mathbf{r}' = 0, \tag{22}$$

where the line integral is now taken along an arbitrary closed curve C. The physical meaning of Eq. (22) is when we move a particle in the gravitational field along an arbitrary closed trajectory, we do not expend any work.

This property must be true also when the closed trajectory (curve) C is infinitesimally short. This means that the gravitational field must be an irrotational vector field: its *vorticity* is equal to 0 everywhere:

$$\nabla \times \mathbf{g}(\mathbf{r}) = \mathbf{0}. \tag{23}$$

A field which behaves in this way is also known as a *potential field*, meaning that there exists a scalar function, called *potential*, of which the vector field in question is a *gradient*. Denoting this potential by $W(\mathbf{r})$, we can thus write

$$\nabla W(\mathbf{r}) = \mathbf{g}(\mathbf{r}). \tag{24}$$

To get some insight into the physical meaning of the potential W, whose physical units are m^2 s^{-2}, we relate it to the work w defined in Eq. (21). It can be shown that the amount of work expended when moving a mass m from $\mathbf{r}_1$ to $\mathbf{r}_2$, along an arbitrary trajectory, is equal to

$$w = [W(\mathbf{r}_2) - W(\mathbf{r}_1)]m. \tag{25}$$

In addition to the two differential equations (Eqs. 23 and 24) governing the behavior of the gravity field, there is a third equation describing the field's *divergence*,

$$\nabla \cdot \mathbf{g}(\mathbf{r}) = -4\pi G\rho(\mathbf{r}) + 2\omega^2. \tag{26}$$

These three *field equations* describe fully the differential behavior of the earth's gravity field. We note that the first term on the right-hand side of Eq. (26) corresponds to the gravitational potential W_g, whose gradient is the gravitational vector $\boldsymbol{g}_g$, while the second term corresponds to the centrifugal potential W_c that gives rise to the centrifugal acceleration vector $\boldsymbol{g}_c$. The negative sign of the first term indicates that, at the point $\mathbf{r}$, there is a sink rather than a source of the gravity field, which should be somewhat obvious from the direction of the vectors of the field. Since the ∇ operator is linear, we can write

$$W(\mathbf{r}) = W_g(\mathbf{r}) + W_c(\mathbf{r}). \tag{27}$$

A potential field is a scalar field that is simple to describe and to work with, and it has become the basic descriptor of the earth's gravity field in geodesy (cf., the article "Global Gravity" in this volume). Once one has an adequate knowledge of the gravity potential, one can derive all the other characteristics of the earth's gravity field, $\mathbf{g}$ by Eq. (24), W_g by Eq. (27), etc., mathematically. Interestingly, the Newton integral in Eq. (16) can be also rewritten for the gravitational potential W_g, rather than the acceleration, to give

$$W_g(\mathbf{r}) = G\int_B \rho(\mathbf{r}')|\mathbf{r}' - \mathbf{r}|^{-1}\, dV, \tag{28}$$

which is one of the most often used equations in gravity field studies. Let us, for completeness, spell out also the equation for the centrifugal potential:

$$W_c(\mathbf{r}) = \tfrac{1}{2}\,\omega^2 p^2(\mathbf{r}), \tag{29}$$

[cf., Eq. (19)].

A surface on which the gravity potential value is constant is called an *equipotential surface*. As the value of the potential varies continuously, we may recognize infinitely many equipotential surfaces defined by the following prescription:

$$W(\mathbf{r}) = \text{const}. \tag{30}$$

These equipotential surfaces are convex everywhere above the earth and never cross each other anywhere. By definition, the equipotential surfaces are horizontal everywhere and are thus sometimes called the level surfaces.

One of these infinitely many equipotential surfaces is the *geoid*, one of the most important surfaces used in geodesy, the equipotential surface defined by a specific value W_0 and thought of as approximating the MSL the best (cf., Section II.B) in some sense. We shall have more to say about the two requirements in Section IV.D. At the time of writing, the best value of W_0 is thought to be $62{,}636{,}855.8 \pm 0.5$ m^2 s^{-2} (Burša *et al.*, 1997). A global picture of the geo id is shown in Fig. 5 in the article "Global Gravity" in this volume, where the geoidal height N (cf., Section II.B), i.e., the geoid-ellipsoid separation, is plotted. Note that the departure of the geoid from the mean earth ellipsoid (for the definition see below) is at most about 100 m in absolute value.

When studying the earth's gravity field, we often need and use an idealized model of the field. The use of such a model allows us to express the actual gravity field as

a sum of the selected model field and the remainder of the actual field. When the model field is selected to resemble closely the actual field itself, the remainder, called an anomaly or disturbance, is much smaller than the actual field itself. This is very advantageous because working with significantly smaller values requires less rigorous mathematical treatment to arrive at the same accuracy of the final results. This procedure resembles the "linearization" procedure used in mathematics and is often referred to as such. Two such models are used in geodesy: spherical (radial field) and ellipsoidal (also called normal or Somigliana-Pizzetti's) models. The former model is used mainly in satellite orbit analysis and prediction (cf., Section V.C), while the normal model is used in terrestrial investigations.

The normal gravity field is generated by a massive body called the *mean earth ellipsoid* adopted by a convention. The most recent such convention, proposed by the IAG in 1980 (IAG, 1980) and called Geodetic Reference System of 1980 (GRS 80), specifies the mean earth ellipsoid as having the major semi-axis "a" 6,378,137 m long and the flattening "f" of 1/298.25. A flattening of an ellipsoid is defined as

$$f = (a - b)/a, \tag{31}$$

where "b" is the minor semi-axis. This ellipsoid departs from a mean earth sphere by slightly more than 10 km; the difference of $a - b$ is about 22 km. We must note here that the flattening f is closely related to the second degree coefficient $C_{2,0}$ discussed in the article "Global Gravity."

This massive ellipsoid is defined as rotating with the earth with the same angular velocity ω, its potential is defined to be constant and equal to W_0 on the surface of the ellipsoid, and its mass is the same as that (M) of the earth. Interestingly, these prescriptions are enough to evaluate the *normal potential* $U(\mathbf{r})$ everywhere above the ellipsoid so that the mass density distribution within the ellipsoid does not have to be specified. The departure of the actual gravity potential from the normal model is called *disturbing potential* $T(\mathbf{r})$:

$$T(\mathbf{r}) = W(\mathbf{r}) - U(\mathbf{r}). \tag{32}$$

The earth's gravity potential field is described in a global form as a truncated series of spherical harmonics up to order and degree 360 or even higher. Many such series have been prepared by different institutions in the United States and in Europe. Neither regional nor local representations of the potential are used in practice; only the geoid, the gravity anomalies, and the deflections of the vertical (see the next three sections) are needed on a regional/local basis.

C. Magnitude of Gravity

The gravity vector $\mathbf{g}(\mathbf{r})$ introduced in Section III.A can be regarded as consisting of a magnitude (length) and a direction. The magnitude of $\mathbf{g}$, denoted by g, is referred to as *gravity*, which is a scalar quantity measured in units of acceleration. It changes from place to place on the surface of the earth in response to latitude, height, and the underground mass density variations. The largest is the latitude variation, due to the oblateness of the earth and due to the change in centrifugal acceleration with latitude, with amounts to about 5.3 cm s^{-2}, i.e., about 0.5% of the total value of gravity. At the poles, gravity is the strongest, about 9.833 m s^{-2} (983.3 Gal); at the equator it is at its weakest, about 978.0 Gal. The height variation, due to varying distance from the attracting body, the earth, amounts to 0.3086 mGal m^{-1} when we are above the earth and to around 0.0848 mGal m^{-1} (we have seen this gradient already in Section II.B) when we are in the uppermost layer of the earth such as the topography. The variations due to mass density variations are somewhat smaller. We note that all these variations in gravity are responsible for the variation in weight: for instance, a mass of 1 kg at the pole weighs 9.833 kg m^1 s^{-2}, while on the equator it weighs only 9.780 kg m^1 s^{-2}.

Gravity can be measured by means of a test mass, by simply measuring either the acceleration of the test mass in free fall or the force needed to keep it in place. Instruments that use the first approach, pendulums and "free-fall devices," can measure the total value of gravity (absolute instruments), while the instruments based on the second approach, called "gravimeters," are used to measure gravity changes from place to place or changes in time (relative instruments). The most popular field instruments are the gravimeters (of many different designs) which can be made easily portable and easily operated. *Gravimetric surveys* conducted for the geophysical exploration purpose, which is the main user of detailed gravity data, employ portable gravimeters exclusively. The accuracy, in terms of standard deviations, of most of the data obtained in the field is of the order of 0.05 mGal.

To facilitate the use of gravimeters as relative instruments, countries have developed *gravimetric networks* consisting of points at which gravity had been determined through a national effort. The idea of gravimetric networks is parallel to the geodetic (positioning) networks we have seen in Section II.D, and the network adjustment process is much the same as the one used for the geodetic networks. A national gravimetric network starts with national gravity reference point(s) established in participating countries through an international effort; the last such effort, organized by IAG (cf., Section I.C), was the International Gravity Standardization Net 1971 (IGSN 71) (IAG, 1974).

Gravity data as observed on the earth's surface are of little direct use in exploration geophysics. To become useful, they have to be stripped of

1. The height effect, by reducing the observed gravity to the geoid, using an appropriate vertical gradient of gravity $\partial g/\partial H$
2. The dominating latitudinal effect, by subtracting from them the corresponding magnitude of normal gravity γ [the magnitude of the gradient of U, cf., Eq. (24)] reckoned on the mean earth ellipsoid (see Section III.B), i.e., at points (r_e, φ, λ)

The resulting values Δg are called *gravity anomalies* and are thought of as corresponding to locations (r_g, φ, λ) on the geoid. For geophysical interpretation, gravity anomalies are thus defined as (for the definition used in geodesy, see Section III.E)

$$\Delta g(r_g, \varphi, \lambda) = g(r_t, \varphi, \lambda) - \partial g/\partial H \; H(\varphi, \lambda) - \gamma(r_e, \varphi, \lambda), \tag{33}$$

where $g(r_t, \varphi, \lambda)$ are the gravity values observed at points (r_t, φ, λ) on the earth's surface and $H(\varphi, \lambda)$ are the orthometric heights of the observed gravity points. These orthometric heights are usually determined together with the observed gravity. Normal gravity on the mean earth ellipsoid is part of the normal model accepted by convention as discussed in the previous section. The GRS 80 specifies the normal gravity on the mean earth ellipsoid by the following formula:

$$\gamma(r_e, \varphi) = 978.0327(1 + 0.0052790414 \sin^2 \varphi + 0.0000232718 \sin^4 \varphi + 0.0000001262 \sin^6 \varphi)\text{Gal}. \tag{34}$$

Gravity anomalies, like the disturbing potential in Section III.B, are thought of as showing only the anomalous part of gravity, i.e., the spatial variations in gravity caused by subsurface mass density variations. Depending on what value is used for the vertical gradient of gravity $\partial g/\partial H$, we get different kinds of gravity anomalies: using $\partial g/\partial H = -0.3086$ mGal m^{-1}, we get the *free-air gravity anomaly*; using $\partial g/\partial H = \frac{1}{2}(-0.3086 - 0.0848)$ mGal m^{-1}, we get the (simple) *Bouguer gravity anomaly*. Other kinds of anomalies exist, but they are not very popular in other than specific theoretical undertakings.

Observed gravity data, different kinds of point gravity anomalies, anomalies averaged over certain geographical cells, and other gravity-related data are nowadays available in either a digital form or in the form of maps. These can be obtained from various national and international agencies upon request. Figure 10 shows the map of free-air gravity anomalies in Canada.

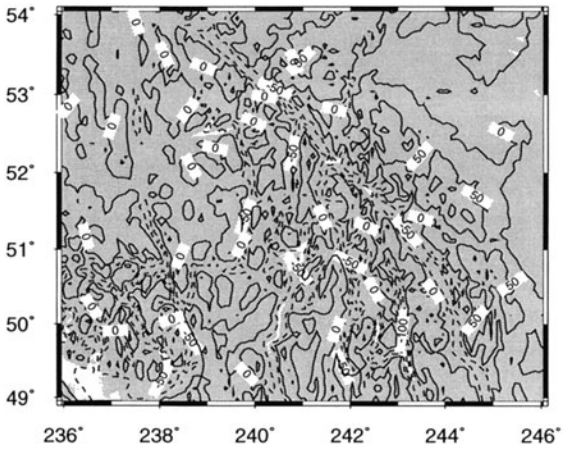


FIGURE 10 Map of free-air gravity anomalies in Canadian Rocky Mountains.

D. Direction of Gravity

Like the magnitude of the gravity vector **g** discussed in the previous section, its direction is also of interest. As it requires two angles to specify the direction, the direction of gravity is a little more difficult to deal with than the magnitude. As has been the case with gravity anomalies, it is convenient to use the normal gravity model here as well. When subtracting the direction of normal gravity from the direction of actual gravity, we end up with a small angle, probably smaller than 1 or 2 arcmin anywhere on earth. This smaller angle θ, called the *deflection of the vertical*, is easier to work with than the arbitrarily large angles used for describing the direction of **g**. We thus have

$$\theta(\mathbf{r}) = \sphericalangle[\mathbf{g}(\mathbf{r}), \boldsymbol{\gamma}(\mathbf{r})], \tag{35}$$

where, in parallel with Eq. (24), $\boldsymbol{\gamma}(\mathbf{r})$ is evaluated as the gradient of the normal potential U:

$$\boldsymbol{\gamma}(\mathbf{r}) = \boldsymbol{\nabla} U(\mathbf{r}). \tag{36}$$

We may again think of the deflection of the vertical as being only just an effect of a disturbance of the actual field, compared to the normal field.

Gravity vectors, being gradients of their respective potentials, are always perpendicular to the level surfaces, be they actual gravity vectors or normal gravity vectors. Thus, the direction of $\mathbf{g}(\mathbf{r})$ is the real vertical (a line perpendicular to the horizontal surface) at $\mathbf{r}$ and the direction of $\boldsymbol{\gamma}(\mathbf{r})$ is the normal vertical at $\mathbf{r}$: the deflection of the

vertical is really the angle between the actual and normal vertical directions. We note that the actual vertical direction is always tangential to the actual *plumbline*, known in physics also as the line of force of the earth's gravity field. At the geoid, for $\mathbf{r} = \mathbf{r}_g$, the direction of $\boldsymbol{\gamma}(\mathbf{r}_g)$ is to a high degree of accuracy the same as the direction of the normal to the mean earth ellipsoid (being exactly the same on the mean earth ellipsoid).

If the mean earth ellipsoid is chosen also as a reference ellipsoid, then the angles that describe the direction of the normal to the ellipsoid are the geodetic latitude $\boldsymbol{\varphi}$ and longitude $\boldsymbol{\lambda}$ (cf., Section II.A). Similarly, the direction of the plumbline at any point $\mathbf{r}$ is defined by *astronomical latitude* $\boldsymbol{\Phi}$ and *astronomical longitude* $\boldsymbol{\Lambda}$. The astronomical coordinates $\boldsymbol{\Phi}$ and $\boldsymbol{\Lambda}$ can be obtained, to a limited accuracy, from optical astronomical measurements, while the geodetic coordinates are obtained by any of the positioning techniques described in Section II. Because θ is a spatial angle, it is customary in geodesy to describe it by two components, the meridian ξ and the prime vertical η components. The former is the projection of θ onto the local meridian plane, and the latter is the projection onto the local prime vertical plane (plane perpendicular to the horizontal and meridian planes).

There are two kinds of deflection of vertical used in geodesy: those taken at the surface of the earth, at points $\mathbf{r}_t = (r_t, \varphi, \lambda)$, called *surface deflections* and those taken at the geoid level, at points $\mathbf{r}_g = (r_g, \varphi, \lambda)$, called *geoid deflections*. Surface deflections are generally significantly larger than the geoid deflections, as they are affected not only by the internal distribution of masses but also by the topographical masses. The two kinds of deflections can be transformed to each other. To do so, we have to evaluate the curvature of the plumbline (in both perpendicular directions) and the curvature of the normal vertical. The former can be quite sizeable—up to a few tens of arcseconds—and is very difficult to evaluate. The latter is curved only in the meridian direction (the normal field being rotationally symmetrical), and even that curvature is rather small, reaching a maximum of about 1 arcsec.

The classical way of obtaining the deflections of the vertical is through the differencing of the astronomical and geodetic coordinates as follows:

$$\xi = \boldsymbol{\Phi} - \varphi, \eta = (\boldsymbol{\Lambda} - \lambda)\cos\varphi. \tag{37}$$

These equations also define the signs of the deflection components. In North America, however, the sign of η is sometimes reversed. We emphasize here that the geodetic coordinates have to refer to the geocentric reference ellipsoid/mean earth ellipsoid. Both geodetic and astronomical coordinates must refer to the same point, either on the geoid or on the surface of the earth. In Section II.B, we mentioned that the astronomical determination of point positions ($\boldsymbol{\Phi}$, $\boldsymbol{\Lambda}$) is not used in practice any more because of the large effect of the earth's gravity field. Here, we see the reason spelled out in Eqs. (37): considering the astronomically determined position ($\boldsymbol{\Phi}$, $\boldsymbol{\Lambda}$) to be an approximation of the geodetic position (φ, λ) invokes an error of (ξ, $\eta/\cos\varphi$) that can reach several kilometers on the surface of the earth. The deflections of the vertical can be determined also from other measurements, which we will show in the next section.

E. Transformations between Field Parameters

Let us begin with the transformation of the geoidal height to the deflection of the vertical [i.e., $N \rightarrow (\xi, \eta)$], which is of a purely geometrical nature and fairly simple. When the deflections are of the "geoid" kind, they can be interpreted simply as showing the slope of the geoid with respect to the geocentric reference ellipsoid at the deflection point. This being the case, geoidal height differences can be constructed from the deflections ($(\xi, \eta) \rightarrow \Delta N$) in the following fashion. We take two adjacent deflection points and project their deflections onto a vertical plane going through the two points. These projected deflections represent the projected slopes of the geoid in the vertical plane; their average multiplied by the distance between the two points gives us an estimate of the difference in geoidal heights ΔN at the two points. Pairs of deflection points can be then strung together to produce the geoid profiles along selected strings of deflection points. This technique is known as *Helmert's levelling*. We note that if the deflections refer to a geodetic datum (rather than to a geocentric reference ellipsoid), this technique gives us geoidal height differences referred to the same geodetic datum. Some older geoid models were produced using this technique.

Another very useful relation (transformation) relates the geoid height N to the disturbing potential T ($T \rightarrow N$, $N \rightarrow T$). It was first formulated by a German physicist H. Bruns (1878), and it reads

$$N = T/\gamma. \tag{38}$$

The equation is accurate to a few millimeters; it is now referred to as *Bruns's formula*.

In Section III.C we introduced gravity anomaly Δg of different kinds (defined on the geoid), as they are normally used in geophysics. In geodesy we need a different gravity anomaly, one that is defined for any location $\mathbf{r}$ rather than being tied to the geoid. Such gravity anomaly is defined by the following exact equation:

$$\Delta g(\mathbf{r}) = -\partial T/\partial h|_{\mathbf{r}=(\mathrm{r},\varphi,\lambda)} + \gamma(\mathbf{r})^{-1}\partial\gamma/\partial h|_{\mathbf{r}=(\mathrm{r},\varphi,\lambda)} T(r - Z, \varphi, \lambda), \tag{39}$$

where Z is the displacement between the actual equipotential surface $W = \text{const.}$ passing through $\mathbf{r}$ and the corresponding (i.e., having the same potential) normal equipotential surface $U = \text{const.}$ This differential equation of first order, sometimes called *fundamental gravimetric equation*, can be regarded as the transformation from $T(\mathbf{r})$ to $\Delta g(\mathbf{r})(T \to \Delta g)$ and is used as such in the studies of the earth's gravity field. The relation between this gravity anomaly and the ones discussed above is somewhat tenuous.

Perhaps the most important transformation is that of gravity anomaly Δg, it being the cheapest data, to disturbing potential $T(\Delta g \to T)$, from which other quantities can be derived using the transformations above. This transformation turns out to be rather complicated: it comes as a solution of a scalar boundary value problem and it will be discussed in the following two sections. We devote two sections to this problem because it is regarded as central to the studies of earth's gravity field. Other transformations between different parameters and quantities related to the gravity field exist, and the interested reader is advised to consult any textbook on geodesy; the classical textbook by Heiskanen and Moritz (1967) is particularly useful.

F. Stokes's Geodetic Boundary Value Problem

The scalar *geodetic boundary value problem* was formulated first by Stokes (1849). The formulation is based on the partial differential equation valid for the gravity potential W [derived by substituting Eq. (24) into Eq. (26)],

$$\nabla^2 W(\mathbf{r}) = -4\pi G\rho(\mathbf{r}) + 2\omega^2. \tag{40}$$

This is a nonhomogeneous elliptical equation of second order, known under the name of *Poisson equation*, that embodies all the field equations (see Section III.A) of the earth gravity field. Stokes applied this to the disturbing potential T (see Section III.B) outside the earth to get

$$\nabla^2 T(\mathbf{r}) = 0. \tag{41}$$

This is so because T does not have the centrifugal component and the mass density $\rho(\mathbf{r})$ is equal to 0 outside the earth. (This would be true only if the earth's atmosphere did not exist; as it does exist, it has to be dealt with. This is done in a corrective fashion, as we shall see below.) This homogeneous form of Poisson equation is known as *Laplace equation*. A function (T, in our case) that satisfies the Laplace equation in a region (outside the earth, in our case) is known as being *harmonic* in that region.

Further, Stokes has chosen the geoid to be the boundary for his boundary value problem because it is a smooth enough surface for the solution to exist (in the space outside the geoid). This of course violates the requirement of harmonicity of T by the presence of topography (and the atmosphere). Helmert (1880) suggested to avoid this problem by transforming the formulation into a space where T is harmonic outside the geoid. The actual disturbing potential T is transformed to a disturbing potential T^{h}, harmonic outside the geoid, by subtracting from it the potential caused by topography (and the atmosphere) and adding to it the potential caused by topography (and the atmosphere) condensed on the geoid (or some other surface below the geoid). Then the Laplace equation

$$\nabla^2 T^{\mathrm{h}}(\mathbf{r}) = 0 \tag{42}$$

is satisfied everywhere outside the geoid. This became known as the *Stokes-Helmert formulation*.

The boundary values on the geoid are constructed from gravity observed on the earth's surface in a series of steps. First, gravity anomalies on the surface are evaluated from Eq. (33) using the free-air gradient These are transformed to Helmert's anomalies Δg^{h}, defined by Eq. (39) for $T = T^{\mathrm{h}}$, by applying a transformation parallel to the one for the disturbing potentials as described above. By adding some fairly small corrections, Helmert's anomalies are transformed to the following expression (Vaníček *et al.*, 1999):

$$2r^{-1}T^{\mathrm{h}}(r - Z, \varphi, \lambda) + \partial T^{\mathrm{h}}/\partial \mathrm{r}|_{\mathrm{r}} = -\Delta g^{\mathrm{h*}}(\mathbf{r}). \tag{43}$$

As T^{h} is harmonic above the geoid, this linear combination, multiplied by r, is also harmonic above the geoid. As such it can be "continued downward" to the geoid by using the standard Poisson integral.

Given the Laplace equation (42), the boundary values on the geoid, and the fact that $T^{\mathrm{h}}(\mathbf{r})$ disappears as $r \to \infty$, Stokes (1849) derived the following integral solution to his boundary value problem:

$$T^{\mathrm{h}}(\mathbf{r}_{\mathrm{g}}) = T^{\mathrm{h}}(r_g, \Omega)R/(4\pi) \int_G \Delta g^{\mathrm{h*}}(r_g, \Omega')\, S(\Psi)\, d\Omega', \tag{44}$$

where Ω, Ω' are the geocentric spatial angles of positions $\mathbf{r}$, $\mathbf{r}'$, Ψ is the spatial angle between $\mathbf{r}$ and $\mathbf{r}'$; S is the Stokes integration kernel in its spherical (approximate) form

$$S(\Psi) \cong 1 + \sin^{-1}(\Psi/2) - 6\sin(\Psi/2) - 5\cos\Psi - 3\cos\Psi \ln[\sin(\Psi/2) + \sin^2(\Psi/2)]; \tag{45}$$

and the integration is carried out over the geoid. We note that, if desired, the disturbing potential is easily transformed to geoidal height by means of the Bruns formula (38).

In the final step, the solution $T^{\mathrm{h}}(\mathbf{r}_{\mathrm{g}})$ is transformed to $T(\mathbf{r}_{\mathrm{g}})$ by adding to it the potential of topography (and the atmosphere) and subtracting the potential of topography (and the atmosphere) condensed to the geoid. This can be regarded as a back transformation from the "Helmert

harmonic space" back to the real space. We have to mention that the fore and back transformation between the two spaces requires knowledge of topography (and the atmosphere), both of height and of density. The latter represents the most serious accuracy limitation of Stokes's solution: the uncertainty in topographical density may cause an error up to 1 to 2 dm in the geoid in high mountains.

Let us add that recently it became very popular to use a higher than second order (Somigliana-Pizzetti's) reference field in Stokes's formulation. For this purpose, a global field (cf., the article, "Global Gravity"), preferably of a pure satellite origin, is selected and a residual disturbing potential on, or geoidal height above, a *reference spheroid* defined by such a field (cf., Fig. 5 in the article "Global Gravity") is then produced. This approach may be termed a *generalized Stokes formulation* (Vaníček and Sjöberg, 1991), and it is attractive because it alleviates the negative impact of the existing nonhomogeneous terrestrial gravity coverage by attenuating the effect of distant data in the Stokes integral (44). For illustration, so computed a geoid for a part of North America is shown in Fig. 11. It should be also mentioned that the evaluation of Stokes' integral is often sought in terms of Fast Fourier Transform.

G. Molodenskij's Geodetic Boundary Value Problem

In the mid-20th century, Russian physicist M. S. Molodenskij formulated a different scalar boundary value problem to solve for the disturbing potential outside the earth (Molodenskij, Eremeev, and Yurkine 1960). His criticism of Stokes' approach was that the geoid is an equipotential surface internal to the earth and as such requires detailed knowledge of internal (topographical) earth mass density, which we will never have. He then proceeded to replace Stokes's choice of the boundary (geoid) by the earth's surface and to solve for $T(\mathbf{r})$ outside the earth.

At the earth's surface, the Poisson equation changes dramatically. The first term on its right-hand side, equal to

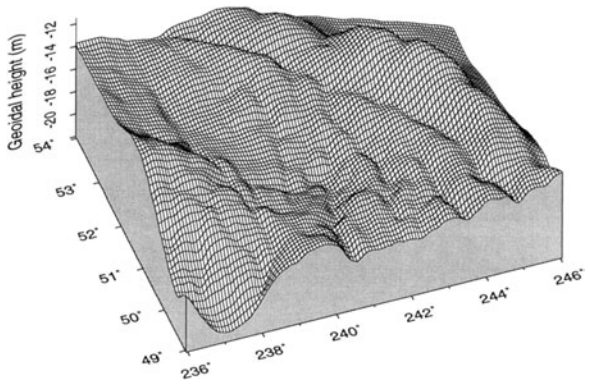


FIGURE 11 Detailed geoid for an area in the Canadian Rocky Mountains (computed at the University of New Brunswick).

$-4\pi G\rho(\mathbf{r})$, changes from 0 to a value of approximately $2.24 * 10^{-6}$ s^{-2} (more than three orders of magnitude larger than the value of the second term). The latter value is obtained using the density ρ of the most common rock, granite. Conventionally, the value of the first term right on the earth's surface is defined as $-4k(\mathbf{r}_t)\pi G\rho(\mathbf{r}_t)$, where the function $k(\mathbf{r}_t)$ has a value between 0 and 1 depending on the shape of the earth's surface; it equals 1/2 for a flat surface, close to 0 for a "needle-like" topographical feature, and close to 1 for a "well-like" feature. In Molodenskij's solution, the Poisson equation has to be integrated over the earth's surface and the above variations of the right-hand side cause problems, particularly on steep surfaces. It is still uncertain just how accurate a solution can be obtained with Molodenskij's approach; it looks as if bypassing the topographical density may have introduced another problem caused by the real shape of topographical surface.

For technical reasons, the integration is not carried out on the earth's surface but on a surface which differs from the earth's surface by about as much as the geoidal height N; this surface is the telluroid encountered already in Section II.B. The solution for T on the earth's surface (more accurately on the telluroid) is given by the following integral equation:

$$\begin{aligned} T(\mathbf{r}_t) - R/(2\pi) \int_{tell} & [\partial/\partial n' |\mathbf{r}_t - \mathbf{r}'_t|^{-1} - |\mathbf{r}_t - \mathbf{r}'_t|^{-1} \\ & \times \cos\beta/\gamma\, \partial\gamma/\partial H^N] T(\mathbf{r}'_t)\, d\Omega' \\ = R/(2\pi) \int_{tell} & [\Delta g(\mathbf{r}'_t) - \gamma][\xi' \tan\beta_1 + \eta' \tan\beta_2] \\ & \times |\mathbf{r}_t - \mathbf{r}'_t|^{-1} \cos\beta\, d\Omega', \end{aligned} \tag{46}$$

where n' is the outer normal to the telluroid; β is the maximum slope of the telluroid (terrain); β_1, β_2 are the north–south and east–west terrain slopes; and ξ', η' are deflection components on the earth's surface. This integral equation is too complicated to be solved directly and simplifications must be introduced. The solution is then sought in terms of successive iterations, the first of which has an identical shape to the Stokes integral (44). Subsequent iterations can be thought of as supplying appropriate corrective terms (related to topography) to the basic Stokes solution.

In fact, the difference between the telluroid and the earth's surface, called the *height anomaly* ζ, is what can be determined directly from Molodenskij's integral using a surface density function. It can be interpreted as a "geoidal height" in Molodenskij's sense as it defines the Molodenskij "geoid" introduced in Section II.B (called quasi-geoid, to distinguish it from the real geoid). The difference between the geoid and quasi-geoid may reach up to a few meters in mountainous regions, but it disappears at sea (Pick, Pícha, and Vyskočil, 1973). It can be seen from

Section II.B that the difference may be evaluated from orthometric and normal heights (referred to the geoid, and quasi-geoid, respectively):

$$\zeta - N = H^{\mathrm{O}} - H^{\mathrm{N}}, \tag{47}$$

subject to the error in the orthometric height. Approximately, the difference is also equal to $-\Delta g^{\mathrm{Bouguer}} H^{\mathrm{O}}/\gamma$.

H. Global and Local Modeling of the Field

Often, it is useful to describe the different parameters of the earth's gravity field by a series of spherical or ellipsoidal harmonic functions (cf., the article "Global Gravity" in this volume). This description is often referred to as the *spectral form*, and it is really the only practical global description of the field parameters. The spectral form, however, is useful also in showing the spectral behavior of the individual parameters. We learn, for instance, that the series for T, N, or ζ converge to 0 much faster than the series for Δg, ζ, η do: we say that the T, N, or ζ fields are smoother than the Δg, ζ, η fields. This means that a truncation of the harmonic series describing one of the smoother fields does not cause as much damage as does a truncation for one of the rougher fields by leaving out the higher "frequency" components of the field. The global description of the smoother fields will be closer to reality.

If higher frequency information is of importance for the area of interest, then it is more appropriate to use a point description of the field. This form of a description is called in geodesy the *spatial form*. Above we have seen only examples of spatial expressions, in the article "Global Gravity" only spectral expressions are used. Spatial expressions involving surface convolution integrals over the whole earth [cf., Eqs. (44) and (46)], can be always transformed into corresponding spectral forms and vice versa. The two kinds of forms can be, of course, also combined as we saw in the case of generalized Stokes's formulation (Section III.F).

IV. GEO-KINEMATICS

A. Geodynamics and Geo-Kinematics

Dynamics is that part of physics that deals with forces (and therefore masses), and motions in response to these forces and geodynamics is that part of geophysics that deals with the dynamics of the earth. In geodesy, the primary interest is the geometry of the motion and of the deformation (really just a special kind of motion) of the earth or its part. The geometrical aspect of dynamics is called kinematics, and therefore, we talk here about *geo-kinematics*. As a matter of fact, the reader might have noticed already in the above paragraphs involving physics how mass was eliminated from the discussions, leaving us with only kinematic descriptions.

Geo-kinematics is one of the obvious fields where cooperation with other sciences, geophysics here, is essential. On the one hand, geometrical information on the deformation of the surface of the earth is of much interest to a geophysicist who studies the forces/stresses responsible for the deformation and the response of the earth to these forces/stresses. On the other hand, it is always helpful to a geodesist to get an insight into the physical processes responsible for the deformation he is trying to monitor.

Geodesists have studied some parts of geo-kinematics, such as those dealing with changes in the earth's rotation, for a long time. Other parts were only more recently incorporated into geodesy because the accuracy of geodetic measurements had not been good enough to see the real-time evolution of deformations occurring on the surface of the earth. Nowadays, geodetic monitoring of crustal motions is probably the fastest developing field of geodesy.

B. Temporal Changes in Coordinate Systems

In Section II.A we encountered a reference to the earth's "spin axis" in the context of geocentric coordinate systems. It is this axis the earth spins around with 366.2564 *sidereal revolutions* (cf., Section III.A), or 365.2564 revolutions with respect to the sun (defining *solar days*) per year. The spin axis of the earth moves with respect to the universe (directions to distant stars, the realization of an inertial coordinate system), undergoing two main motions: one very large, called precession, with a period of about 26,000 years and the other much smaller, called nutation, with the main period of 18.6 years. These motions must be accounted for when doing astronomical measurements of either the optical or radio variety (see Section II.C).

In addition to precession and nutation, the spin axis also undergoes a torque-free nutation, also called a wobble, with respect to the earth. More accurately, the *wobble* should be viewed as the motion of the earth with respect to the instantaneous spin axis. It is governed by the famous Euler's gyroscopic equation:

$$\mathbf{J}\,\boldsymbol{\omega} + \boldsymbol{\omega} \times \mathbf{J}\,\boldsymbol{\omega}, \tag{48}$$

where $\mathbf{J}$ is the earth's *tensor of inertia* and $\boldsymbol{\omega}$ is the instantaneous spin *angular velocity vector* whose magnitude ω we have met several times above. Some relations among the diagonal elements of $\mathbf{J}$, known as moments of inertia, can be inferred from astronomical observations giving a solution $\boldsymbol{\omega}$ of this differential equation. Such a solution describes a periodic motion with a period of about 305 sidereal days, called Euler's period.

Observations of the wobble (Fig. 12) have shown that beside the Euler component, there is also an annual periodic component of similar magnitude plus a small drift. The magnitude of the periodic components fluctuates

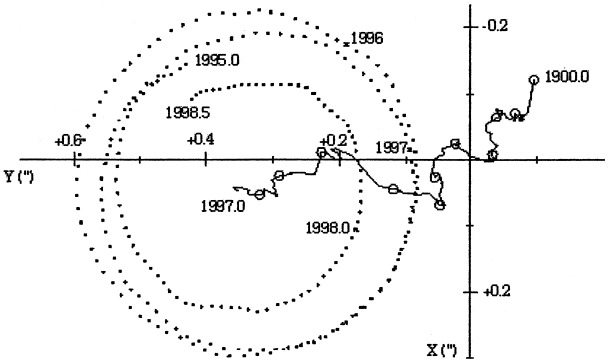


FIGURE 12 Earth pole wobble. (Source: International Earth Rotation Service.)

around 0.1 arcsec. Thus, the wobble causes a displacement of the pole (intersection of instantaneous spin axis with the earth's surface) of several meters. Furthermore, systematic observations show also that the period of the Euler component is actually longer by some 40% than predicted by the Euler equation. This discrepancy is caused by the nonrigidity of the earth, and the actual period, around 435 solar days, is now called Chandler's. The actual motion of the pole is now being observed and monitored by IAG's IERS on a daily basis. It is easy to appreciate that any coordinate system linked to the earth's spin axis (cf., Section II.A) is directly affected by the earth pole wobble which, therefore, has to be accounted for.

As the direction of ω varies, so does its magnitude ω. The variations of spin velocity are also monitored in terms of the *length of the day* (LOD) by the Bureau Internationale des Poids et Mesures—Time Section (BIPM) on a continuous basis. The variations are somewhat irregular and amount to about 0.25 msec/year—the earth's spin is generally slowing down.

There is one more temporal effect on a geodetic coordinate system, that on the datum for vertical positioning, i.e., on the geoid. It should be fairly obvious that if the reference surface for heights (orthometric and dynamic) changes, so do the heights referred to it. In most countries, however, these changes are not taken too seriously. The geoid indicated by the MSL, as described in Section III.B, changes with time both in response to the mass changes within the earth and to the MSL temporal changes. The latter was discussed in Section III.B, the former will be discussed in Section IV.C.

C. Temporal Changes in Positions

The earth is a deformable body and its shape is continuously undergoing changes caused by a host of stresses; thus, the positions of points on the earth's surface change continuously. Some of the stresses that cause the *deformations* are known, some are not, with the best known being the tidal stress (Melchior, 1966). Some loading stresses causing crustal deformation are reasonably well known, such as those caused by filling up water dams, others, such as sedimentation and glaciation, are known only approximately. Tectonic and other stresses can be only inferred from observed deformations (cf., the article, "Tectonophysics" in this encyclopedia). The response of the earth to a stress, i.e., the deformation, varies with the temporal frequency and the spatial extent of the stress and depends on the rheological properties of the whole or just a part of the earth. Some of these properties are now reasonably well known, some are not. The ultimate role of geodesy vis-à-vis these deformations is to take them into account for predicting the temporal variations of positions on the earth's surface. This can be done relatively simply for deformations that can be modeled with a sufficient degree of accuracy (e.g., tidal deformations), but it cannot yet be done for other kinds of deformations, where the physical models are not known with a sufficient degree of certainty. Then the role of geodesy is confined to monitoring the surface movements, kinematics, to provide the input to geophysical investigations.

A few words are now in order about tidal deformations. They are caused by the moon and sun gravitational attraction (see Fig. 13). Tidal potential caused by the next most influential celestial body, Venus, amounts to only 0.036% of the luni-solar potential; in practice only the *luni-solar tidal potential* is considered. The tidal potential $W_t^{\mathbb{C}}$ of the moon, the lunar tidal potential, is given by the following equation:

$$W_t^{\mathbb{C}}(\mathbf{r}) = GM^{\mathbb{C}}|\mathbf{r}^{\mathbb{C}} - \mathbf{r}|^{-1} \sum_{j=2}^{\infty} \mathbf{r}^j |\mathbf{r}^{\mathbb{C}} - \mathbf{r}|^{-j} P_j(\cos \Psi^{\mathbb{C}}), \tag{49}$$

where the symbol $\mathbb{C}$ refers to the moon, P_j is the Legendre function of degree j, and Ψ is the geocentric angle

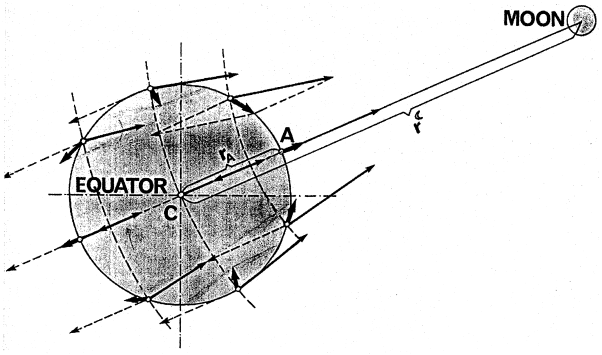


FIGURE 13 The provenance of tidal force due to the moon.

between $\mathbf{r}$ and $\mathbf{r}^c$. The tidal potential of the sun $W_t^\bullet$ is given by a similar series, and it amounts to about 46% of the lunar tidal potential. To achieve an accuracy of 0.03%, it is enough to take the first two terms from the lunar series and the first term from the solar series.

The temporal behavior of tidal potential is periodic, as dictated by the motions of the moon and the sun (but see Section IV.D for the tidal constant effect). The tidal waves, into which the potential can be decomposed, have periods around 1 day, called diurnal; around 12 h, called semidiurnal; and longer. The tidal deformation as well as all the tidal effects (except for the sea tide, which requires solving a boundary value problem for the Laplace tidal equation, which we are not going to discuss here) can be relatively easily evaluated from the tidal potential. This is because the rheological properties of the earth for global stresses and tidal periods are reasonably well known. Geodetic observations as well as positions are affected by tidal deformations, and these effects are routinely corrected for in levelling, VLBI (cf., Section II.C), satellite positioning (cf., Section V.B), and other precise geodetic works. For illustration, the range of orthometric height tidal variation due to the moon is 36 cm, that due to the sun is another 17 cm. With tidal deformation being of a global character, however, these tidal variations are all but imperceptible locally.

Tectonic stresses are not well known, but horizontal motion of tectonic plates has been inferred from various kinds of observations, including geodetic, with some degree of certainty, and different maps of these motions have been published (cf., the article, "Tectonophysics" in this encyclopedia). The AM0-2 absolute plate motion model was chosen to be an "associated velocity model" in the definition of the ITRF (see Section II.A), which together with direct geodetic determination of horizontal velocities define the temporal evolution of the ITRF and thus the temporal evolution of horizontal positions. From other ongoing earth deformations, the post-glacial rebound is probably the most important globally as it is large enough to affect the flattening [Eq. (31)] of the mean earth ellipsoid as we will see in Section IV.C.

Mapping and monitoring of ongoing motions (deformations) on the surface of the earth are done by repeated position determination. In global monitoring the global techniques of VLBI and satellite positioning are used (see Section V.B). For instance, one of the IAG services, the IGS (cf., Section I.C), has been mandated with monitoring the horizontal velocities of a multitude of permanent tracking stations under its jurisdiction. In regional investigations the standard terrestrial geodetic techniques such as horizontal and vertical profiles, horizontal and levelling network re-observation campaigns are employed. The positions are then determined separately from each campaign with subsequent evaluation of *displacements*. Preferably, the displacements (horizontal or vertical) are estimated directly from the observations collected in all the campaigns. The latter approach allows the inclusion of correlations in the mathematical model for the displacement estimation with more correct estimates ensuing. For illustration, Fig. 14 shows such estimated horizontal displacements from the area of Imperial Valley, CA, computed from data covering the period 1941–1975.

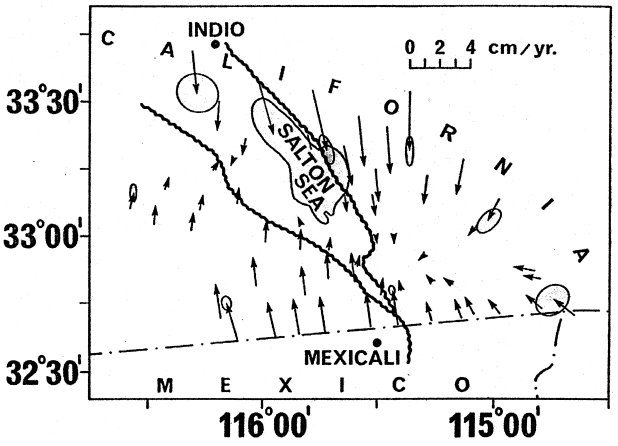


FIGURE 14 Mean annual horizontal displacements in Imperial Valley, CA, computed from data covering the period 1941–1975.

It is not possible to derive absolute displacements from relative positions. Because the repeated horizontal position determination described above is usually of a relative kind, the displacements are indeterminate. It then makes sense to deal only with relative quantification of deformation such as *strain*. Strain is, most generally, described by the displacement gradient matrix $\mathbf{S}$; denoting the 2D displacement vector of a point $\mathbf{r}$ by $\mathbf{v}(\mathbf{r})$ we can write

$$\mathbf{S}(\mathbf{r}) = \nabla' \mathbf{v}^{\mathrm{T}}(\mathbf{r}), \tag{50}$$

where ∇' is the 2D nabla operator. The inverse transformation

$$\mathbf{v}(\mathbf{r}) = \mathbf{S}(\mathbf{r})\,\mathbf{r} + \mathbf{v}_0, \tag{51}$$

where $\mathbf{v}_0$ describes the translational indeterminacy, shows better the role of $\mathbf{S}$, which can be also understood as a Jacobian matrix [cf., Eq. (14)] which transforms from the space of positions (real 2D space) into the space of displacements. The symmetrical part of $\mathbf{S}$ is called the deformation tensor in the mechanics of continuum; the antisymmetrical part of $\mathbf{S}$ describes the rotational deformation. Other strain parameters can be derived from $\mathbf{S}$.

D. Temporal Changes in Gravity Field

Let us begin with the two requirements defining the geoid presented in Section III.B: the constancy of W_0 and the fit of the equipotential surface to the MSL. These two requirements are not compatible when viewed from the point of time evolution or the temporal changes of the earth's gravity field. The MSL grows with time at a rate estimated to be between 1 and 2 mm year^{-1} (the eustatic water level change), which would require systematic lowering of the value of W_0. The mass density distribution within the earth changes with time as well (due to tectonic motions, post-glacial rebound, sedimentation, as discussed above), but its temporal effect on the geoid is clearly different from that of the MSL. This dichotomy has not been addressed by the geodetic community yet.

In the areas of largest documented changes of the mass distribution (those caused by the post-glacial rebound), the northeastern part of North America and Fennoscandia, the maximum earth surface uplift reaches about 1 cm year^{-1}. The corresponding change in gravity value reaches up to 0.006 mGal year^{-1} and the change in the equipotential surfaces $W = \text{const.}$ reaches up to 1 mm year^{-1}. The potential coefficient $C_{2,0}$ (cf., Section III.B), or rather its unitless version $J_{2,0}$ that is used most of the time in gravity field studies, as observed by satellites shows a temporal change caused by the rebound. The rebound can be thought of as changing the shape of the geoid, and thus the shape of the mean earth ellipsoid, within the realm of observability.

As mentioned in Section IV.C, the tidal effect on gravity is routinely evaluated and corrected for in precise gravimetric work, where the effect is well above the observational noise level. By correcting the gravity observations for the periodic tidal variations we eliminate the temporal variations, but we do not eliminate the whole tidal effect. In fact, the luni-solar tidal potential given by Eq. (49) has a significant constant component responsible for what is called in geodesy *permanent tide*. The effect of permanent tide is an increased flattening of gravity equipotential surfaces, and thus of the mean earth ellipsoid, by about one part in 10^5. For some geodetic work, the tideless mean earth ellipsoid is better suited than the mean tide ellipsoid, and, consequently, both ellipsoids can be encountered in geodesy.

Temporal variations of gravity are routinely monitored in different parts of the world. These variations (corrected for the effect of underground water fluctuations) represent an excellent indicator that a geodynamical phenomenon, such as tectonic plate motion, sedimentation loading, volcanic activity, etc., is at work in the monitored region. When gravity monitoring is supplemented with vertical motion monitoring, the combined results can be used to infer the physical causes of the monitored phenomenon. Let us mention that the earth pole wobble introduces also observable variations of gravity of the order of about 0.008 mGal. So far, these variations have been of academic interest only.

V. SATELLITE TECHNIQUES

A. Satellite Motion, Functions, and Sensors

Artificial satellites of the earth appeared on the world scene in the late 1950s and were relatively early embraced by geodesists as the obvious potential tool to solve worldwide geodetic problems. In geodetic applications, satellites can be used both in positioning and in gravitational field studies as we have alluded to in the previous three sections. Geodesists have used many different satellites in the past 40 years, ranging from completely passive to highly sophisticated active (transmitting) satellites, from quite small to very large. Passive satellites do not have any sensors on board and their role is basically that of an orbiting target. Active satellites may carry a large assortment of sensors, ranging from accurate clocks through various counters to sophisticated data processors, and transmit the collected data down to the earth either continuously or intermittently.

Satellites orbit the earth following a trajectory which resembles the Keplerian ellipse that describes the motion in radial field (cf., Section III.A); the higher the satellite is, the closer its orbit to the Keplerian ellipse. Lower orbiting satellites are more affected by the irregularities of the earth's gravitational field, and their *orbit* becomes more perturbed compared to the Keplerian orbit. This curious behavior is caused by the inherent property of gravitational field known as the attenuation of shorter wavelengths of the field with height and can be gleaned from Eq. (4) in the article "Global Gravity." The ratio a/r is always smaller than 1 and thus tends to disappear the faster the larger its exponent l which stands for the spatial wave number of the field. We can see that the *attenuation factor* $(a/r)^{\ell}$ goes to 0 for growing ℓ and growing r; for $r > a$, we have:

$$\lim_{l \to \infty} (a/r)^{\ell} = 0. \tag{52}$$

We shall see in Section V.C how this behavior is used in studying the gravitational field by means of satellites.

Satellite orbits are classified as high and low orbits, polar orbits (when the orbital plane contains the spin axis of the earth), equatorial orbits (orbital plane coincides with the equatorial plane of the earth), and pro-grade and retro-grade or bits (the direction of satellite motion is either eastward or westward). The lower the orbit is, the faster the satellite circles the earth. At an altitude of about 36,000 km, the orbital velocity matches that of the earth's

spin, its orbital period becomes 24 h long, and the satellite moves only in one meridian plane (its motion is neither pro- nor retro-grade). If its orbit is equatorial, the satellite remains in one position above the equator. Such an orbit (satellite) is called *geostationary*.

Satellites are *tracked* from points on the earth or by other satellites using electromagnetic waves of frequencies that can penetrate the ionosphere. These frequencies propagate along a more or less straight line and thus require intervisibility between the satellite and the tracking device (transmitter, receiver, or reflector); they range from microwave to visible (from 30 to 10^9 MHz). The single or double passage of the electromagnetic signal between the satellite and the tracking device is accurately timed and the distance is obtained by multiplying the time of passage by the propagation speed. The propagation speed is close to the speed of light in vacuum, i.e., 299, 792, 460 m s^{-1}, with the departure being due to the delay of the wave passing through the atmosphere and ionosphere.

Tracked satellite orbits are then computed from the measured (observed) distances and known positions of the tracking stations by solving the equations of motion in the earth's gravity field. This can be done quite accurately (to a centimeter or so) for smaller, spherical, homogeneous, and high-flying spacecraft that can be tracked by lasers. Other satellites present more of a problem; consequently, their orbits are less well known. When orbits are extrapolated into the future, this task becomes known as the *orbit prediction*. Orbit computation and prediction are specialized tasks conducted only by larger geodetic institutions.

B. Satellite Positioning

The first satellite used for positioning was a large, light, passive balloon (ECHO I, launched in 1960) whose only role was to serve as a naturally illuminated moving target on the sky. This target was photographed against the star background from several stations on the earth, and directions to the satellite were then derived from the known directions of surrounding stars. From these directions and from a few measured interstation distances, positions of the camera stations were then computed. These positions were not very accurate though because the directions were burdened by large unpredictable refraction errors (cf., Section II.B). It became clear that *range (distance) measurement* would be a better way to go.

The first practical satellite positioning system (TRANSIT) was originally conceived for relatively inaccurate naval navigation and only later was adapted for much more accurate geodetic positioning needs. It was launched in 1963 and was made available for civilian use 4 years later. The system consisted of several active satellites in circular polar orbits of an altitude of 1074 km and an orbital period of 107 min, the positions (ephemeredes given in the OR coordinate system—cf., Section II.A) of which were continuously broadcast by the satellites to the users. Also transmitted by these satellites were two signals at fairly stable frequencies of 150 and 400 MHz controlled by crystal oscillators. The user would then receive both signals (as well as the ephemeris messages) in his specially constructed TRANSIT *satellite receiver* and compare them with internally generated stable signals of the same frequencies. The beat frequencies would then be converted to *range rates* by means of the Doppler equation

$$\lambda_R = \lambda_T(1 + v/c)(1 + v^2/c^2)^{\frac{1}{2}}, \tag{53}$$

where v is the projection of the range rate onto the receiver-satellite direction; λ_R, λ_T are the wavelengths of the received and the transmitted signals; and c is the speed of light in vacuum. Finally, the range rates and the satellite positions computed from the broadcast ephemeredes would be used to compute the generic position $\mathbf{r}$ of the receiver (more accurately, the position of receiver's antenna) in the CT coordinate system (cf., Section II.A). More precise satellite positions than those broadcast by the satellites themselves were available from the U.S. naval ground control station some time after the observations have taken place. This control station would also predict the satellite orbits and upload these predicted orbits periodically into the satellite memories.

At most, one TRANSIT satellite would be always "visible" to a terrestrial receiver. Consequently, it was not possible to determine the sequence of positions (trajectory) of a moving receiver with this system; only *position lines* (lines on which the unknown position would lie) were determinable. For determining an accurate position (1 m with broadcast and 0.2 m with precise ephemeredes) of a stationary point, the receiver would have to operate at that point for several days.

Further accuracy improvement was experienced when two or more receivers were used simultaneously at two or more stationary points, and relative positions in terms of interstation vectors $\Delta\mathbf{r}$ were produced. The reason for the increased accuracy was the attenuation of the effect of common errors/biases (atmospheric delays, orbital errors, etc.) through differencing. This relative or differential mode of using the system became very popular and remains the staple mode for geodetic positioning even with the more modern GPS used today.

In the late 1970s, the U.S. military started experimenting with the GPS (originally called NAVSTAR). It should be mentioned that the military have always been vitally interested in positions, instantaneous and otherwise, and so many developments in geodesy are owed to military

initiatives. The original idea was somewhat similar to that of the now defunct TRANSIT system (active satellites with oscillators on board that transmit their own ephemeredes) but to have several satellites orbiting the earth so that at least four of them would be always "visible" from any point on the earth. Four is the minimal number needed to get an instantaneous 3D position by measuring simultaneously the four ranges to the visible satellites: three for the three coordinates and one for determining the ever-changing offset between the satellite and receiver oscillators.

Currently, there are 28 active GPS satellites orbiting the earth at an altitude of 20,000 km spaced equidistantly in orbital planes inclined 60 arc-degrees with respect to the equatorial plane. Their orbital period is 12 h. They transmit two highly coherent cross-polarized signals at frequencies of 1227.60 and 1575.42 MHz, generated by atomic oscillators (cesium and rubidium) on board, as well as their own (broadcast) ephemeredes. Two pseudo-random timing sequences of frequencies 1.023 and 10.23 MHz—one called P-code for restricted users only and the other called C/A-code meant for general use—are modulated on the two carriers. The original intent was to use the timing codes for observing the ranges for determining instantaneous positions. For geodetic applications, so determined ranges are too coarse and it is necessary to employ the carriers themselves.

Nowadays, there is a multitude of GPS receivers available off the shelf, ranging from very accurate, bulky, and relatively expensive "*geodetic receivers*" all the way to hand-held and wrist-mounted cheap receivers. The cheapest receivers use the C/A-code ranging (to several satellites) in a point-positioning mode capable of delivering an accuracy of tens of meters. At the other end of the receiver list, the most sophisticated geodetic receivers use both carriers for the ranging in the differential mode. They achieve an accuracy of the interstation vector between a few millimeters for shorter distances and better than $S10^{-7}$, where S stands for the interstation distance, for distances up to a few thousand kilometers.

In addition to the global network of tracking stations maintained by the IGS (cf., Section I.C), there have been networks of continually tracking GPS stations established in many countries and regions; for an illustration, see Fig. 15. The idea is that the tracking stations are used as traditional position control stations and the tracking data are as well used for GPS satellite orbit improvement. The stations also provide "*differential corrections*" for roving GPS users in the vicinity of these stations. These corrections are used to eliminate most of the biases (atmospheric delays, orbital errors, etc.) when added to point positions of roving receivers. As a result of the technological and logistical improvements during the past 20 years, GPS positioning is now cheap, accurate, and used almost

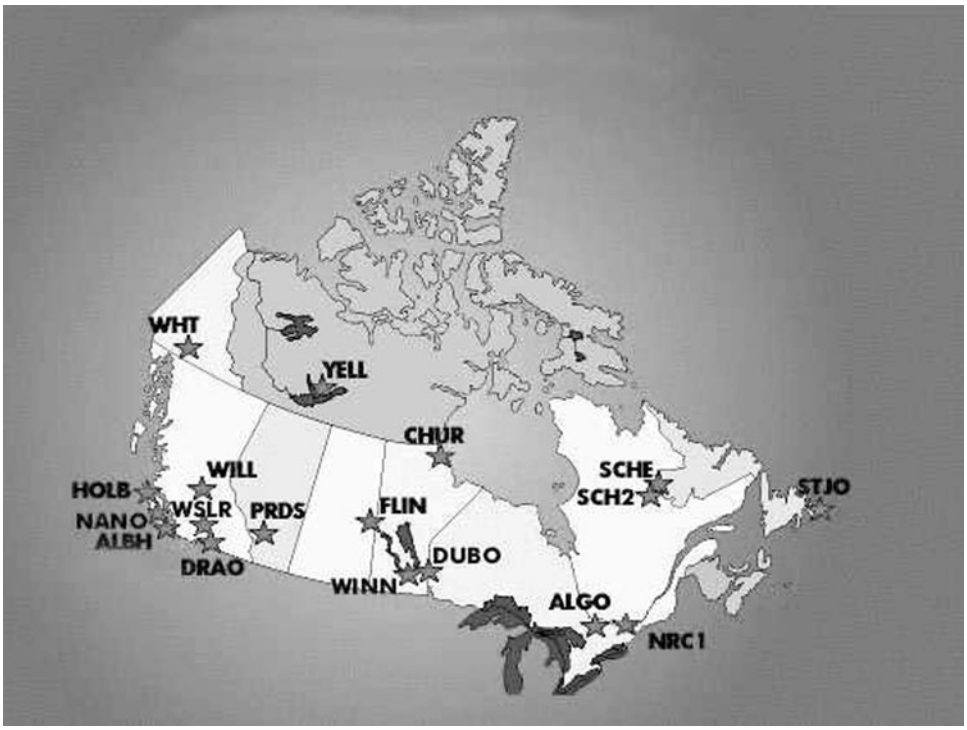


FIGURE 15 Canadian Active Control System. (Source: www.nrcan.gc.ca. Copyright: Her Majesty the Queen in Right of Canada, Natural Resources Canada, Geodetic Survey Division. All rights reserved.)

everywhere for both positioning and precise navigation in preference to classical terrestrial techniques.

The most accurate absolute positions **r** (standard error of 1 cm) are now determined using small, heavy, spherical, high-orbiting, passive satellites equipped with retro-reflectors (LAGEOS 1, LAGEOS 2, STARLETTE, AJISAI, etc.) and laser ranging. The technique became known as *Satellite Laser Ranging* (SLR), and the reason for its phenomenal accuracy is that the orbits of such satellites can be computed very accurately (cf., Section V.A). Also, the ranging is conducted over long periods of time by means of powerful astronomical telescopes and very precise timing devices. Let us just mention that SLR is also used in the relative positioning mode, where it gives very accurate results. The technique is, however, much more expensive than, say, GPS and is thus employed only for scientific investigations

Finally, we have to mention that other satellite-based positioning exist. These are less accurate systems used for nongeodetic applications. Some of them are used solely in commercial application. At least one technique deserves to be pointed out, however, even though it is not a positioning technique per se. This is the synthetic aperture radar interferometry (INSAR). This technique uses collected reflections from a space-borne radar. By sophisticated computer processing of reflections collected during two overflights of the area of interest, the pattern of ground deformation that had occurred between the two flights can be discerned (Massonnet *et al*., 1993). The result is a map of relative local deformations, which may be used, for instance, as a source of information on co-seismic activity. Features as small as a hundred meters across and a decimeter high can be recognized.

C. Gravitational Field Studies by Satellites

The structure of the earth's gravity field was very briefly mentioned in Section III.H, where the field wavelengths were discussed in the context of the spectral description of the global field. A closer look at the field reveals that:

1. The field is overwhelmingly radial (cf., Section III.A) and the first term in the potential series, GM/r, is already a fairly accurate (to about 10^{-3}) description of the field; this is why the radial field is used as a model field (cf., Section III.B) in satellite studies.
2. The largest departure from radiality is described by the second degree term $J_{2,0}$ (cf., Section IV.D), showing the ellipticity of the field, which is about 3 orders of magnitude smaller than the radial part of the field.
3. The remaining wavelength amplitudes are again about 3 orders of magnitude smaller and they further decrease with increasing wave number ℓ. The decrease of amplitude is seen, for instance in Fig. 8 in the article "Global Gravity." In some studies it is possible to use a mathematical expression describing the decrease, such as the experimental Kaula's rule of thumb, approximately valid for ℓ between 2 and 40:

$$\sqrt{\Sigma_{m=2}^{\ell}\left(C_{lm}^2 + S_{lm}^2\right)} \approx R * 10^{-5}/\ell^2, \qquad (54)$$

where C_{lm} and S_{lm} are the potential coefficients (cf., Eq. (4) in "Global Gravity").

As discussed in Section V.A, the earth's gravity field also gets smoother with altitude. Thus, for example, at the altitude of lunar orbit (about 60 times the radius of the earth), the only measurable departure from radiality is due to the earth's ellipticity and even this amounts to less than 3×10^{-7} of the radial component. Contributions of shorter wavelength are 5 orders of magnitude smaller still. Consequently, a low-orbiting satellite has a "bumpier to use a satellite as a gravitation-sensing device, we get more detailed information from low-orbiting spacecraft.

The idea of using satellites to "measure" gravitational field (we note that a satellite cannot sense the total gravity field, cf., Section III.A) stems from the fact that their orbital motion (free fall) is controlled predominantly by the earth's gravitational field. There are other forces acting on an orbiting satellite, such as the attraction of other celestial bodies, air friction, and solar radiation pressure, which have to be accounted for mathematically. Leaving these forces alone, the equations of motion are formulated so that they contain the gravitational field described by potential coefficients C_{lm} and S_{lm}. When the observed orbit does not match the orbit computed from the known potential coefficients, more realistic potential coefficient values can be derived. In order to derive a complete set of more realistic potential coefficients, the procedure has to be formulated for a multitude of different orbits, from low to high, with different inclinations, so that these orbits sample the space above the earth in a homogeneous way. We note that because of the smaller amplitude and faster attenuation of shorter wavelength features, it is possible to use the described *orbital analysis* technique only for the first few tens of degrees ℓ. The article "Global Gravity" shows some numerical results arising from the application of this technique.

Other techniques such as "satellite-to-satellite tracking" and "gradiometry" (see "Global Gravity") are now being used to study the shorter wavelength features of the gravitational field. A very successful technique, "satellite altimetry," a hybrid between a positioning technique and gravitational field study technique (see "Global Gravity") must be also mentioned here. This technique has now been used for some 20 years and has yielded some important results.

SEE ALSO THE FOLLOWING ARTICLES

EARTH SCIENCES, HISTORY OF ● EARTH'S MANTLE (GEOPHYSICS) ● EXPLORATION GEOPHYSICS ● GEOMAGNETISM ● GLOBAL GRAVITY MODELING ● GRAVITATIONAL WAVE ASTRONOMY ● RADIO ASTRONOMY, PLANETARY ● REMOTE SENSING FROM SATELLITES ● TECTONOPHYSICS

BIBLIOGRAPHY

Berthon, S., and Robinson, A. (1991). "The Shape of the World," George Philip Ltd., London.

Boal, J. D., and Henderson, J. P. (1988). The NAD 83 Project—Status and Background. *In* "Papers for the CISM Seminars on the NAD'83 Redefinition in Canada and the Impact on Users" (J. R. Adams ed.), The Canadian Institute of Surveying and Mapping, Ottawa, Canada.

Bomford, G. (1971). "Geodesy," 3rd ed., Oxford Univ. Press, London.

Boucher, C., and Altamini, Z. (1996). "International Terrestrial Reference Frame." *GPS World* **7**(9), 71–75.

Bruns, H. (1878). "Die Figur der Erde," Publication des Königlichen Preussischen Geodätischen Institutes, Berlin, Germany.

Burša, M., Raděj, K., Šíma, Z., True, S. A., and Vatrt, V. (1997). "Determination of Geopotential Scale from Multi-Mission Satellite Altimetry," *Studia Geophys. Geod.* **41,** 203–216.

Grafarend, E. W., and Sansò, F., ed. (1985). "Optimization and Design of Geodetic Networks," Springer-Verlag, Berlin/New York.

Heiskanen, W. A., and Moritz, H. (1967). "Physical Geodesy," Freeman, San Francisco.

Helmert, F. R. (1880). "Die mathematischen und physikalishen Theorien der höheren Geodäsie," Vol. I, Minerva, G. M. B. H. Reprint, 1962.

Hirvonen, R. A. (1971). "Adjustment by Least Squares in Geodesy and Photogrammetry," Ungar, New York.

International Association of Geodesy (1974). "The International Gravity Standardization Net, 1971," Special Publication No. 4, Paris, France.

International Association of Geodesy (1980). "The geodesist's handbook," *Bull. Géodesique* **54**(3).

Lee, L. P. (1976). "Conformal Projections Based on Elliptic," Cartographica Monograph 16, B. V. Gutsell, Toronto, Canada.

Massonnet, D., Rossi, M., Carmona, C., Adragna, F., Peltzer, G., Feigl, K., and Rabaute, T. (1993), "The displacement field of the Landers earthquake mapped by radar interferometry," *Nature* **364**.

Melchior, P. (1966). "The Earth Tides," Pergamon, Elmsford, NY.

Mikhail, E. M. (1976). "Observations and Least Squares," IEP-A dun-Donnelley Publisher.

Molodenskij, M. S., Eremeev, V. F., and Yurkina, M. I. (1960). "Methods for Study of the External Gravitational Field and Figure of the Earth," Translated from Russian by the Israel Program for Scientific Translations for the Office of Technical Services, U.S. Department of Commerce, Washington, DC., 1962.

Mueller, I. I. (1969). "Spherical and Practical Astronomy as Applied to Geodesy," Ungar, New York.

Pick, M., Pícha, J., and Vyskočil, V. (1973). "Theory of the Earth's Gravity Field," Elsevier, Amsterdam/New York.

Ries, J. C., Eanes, R. J., Shum, C. K., and Watkins, M. M. (1992). "Progress in the determination of the gravitational coefficient of the earth," *Geophys. Res. Lett.* **19**(6), 529–531.

Schwarz, C. R., ed. (1989). "North American Datum of 1983," NOAA Professional Paper NOS 2, National Geodetic Information Center, National Oceanic and Atmospheric Administration, Rockville, MD.

Stokes, G. G. (1849). "On the variation of gravity at the surface of the earth," *Trans. Cambridge Philos. Soc.* **VIII,** 672–695.

U.S. Department of Commerce (1973). "The North American Datum," Publication of the National Ocean Survey of NOAA, Rockville, MD.

Vaníček, P., and Krakiwsky, E. J. (1986). "Geodesy: The Concepts," 2nd ed., North-Holland, Amsterdam.

Vaníček, P., and Sjöberg, L. E. (1991). "Reformulation of Stokes's theory for higher than second-degree reference field and modification of integration kernels," *J. Geophys. Res.* **96**(B4), 6529–6539.

Vaníček, P., Huang, J., Novák, P., Véronneau, M., Pagiatakis, S., Martinec, Z., and Featherstone, W. E. (1999). "Determination of boundary values for the Stokes-Helmert problem," *J. Geod.* **73,** pp. 180–192.

Zakatov, P. S. (1953). "A Course in Higher Geodesy," Translated from Russian by the Israel Program form Scientific Translations for the Office of Technical Services, U.S. Department of Commerce, Washington, DC., 1962.

Geostatistics

Clayton V. Deutsch
University of Alberta, Edmonton

GLOSSARY

Declustering Technique to assign relative weights to different data values based on their redundancy with nearby data. Closely spaced data get less weight.
Kriging After the name of D. G. Krige, this term refers to the procedure of constructing the best linear unbiased estimate of a value at a point or of an average over a volume.
Realization Nonunique grid of simulated values. A set of realizations is used as a measure of uncertainty in the variable being studied.
Simulation Procedure of adding correlated error by Monte Carlo to create a value that reflects the full variability.
Variogram Basic tool of the theory used to characterize the spatial continuity of the variable.

GEOSTATICS commonly refers to the theory of regionalized variables and the related techniques that are used to predict variables such as rock properties at unsampled locations. Matheron formalized this theory in the early 1960s (Matheron, 1971). Geostatistics was not developed as a theory in search of practical problems. On the contrary, development was driven by engineers and geologists faced with real problems. They were searching for a consistent set of numerical tools that would help them address real problems such as ore reserve estimation, reservoir performance forecasting, and environmental site characterization. Reasons for seeking such comprehensive technology included (1) an increasing number of data to deal with, (2) a greater diversity of available data at different scales and levels of precision, (3) a need to address problems with consistent and reproducible methods, (4) a belief that improved numerical models should be possible by exploiting computational and mathematical developments in related scientific disciplines, and (5) a belief that more responsible decisions would be made with improved numerical models. These reasons explain the continued expansion of the theory and practice of geostatistics. Problems in mining, such as unbiased estimation of recoverable reserves, initially drove the development of geostatistics. Problems in petroleum, such as realistic heterogeneity models for unbiased flow predictions, were dominant from the mid-1980s through the late 1990s.

Geostatistics is applied extensively in these two areas and is increasingly applied to problems of spatial modeling and uncertainty in environmental studies, hydrogeology, and agriculture.

I. ESSENTIAL CONCEPTS

Geostatistics is concerned with constructing high-resolution three-dimensional models of categorical variables such as rock type or facies and continuous variables such as mineral grade, porosity, or contaminant concentration. It is necessary to have *hard* truth measurements at some volumetric scale. All other data types, including remotely sensed data, are called *soft* data and must be calibrated to the hard data. It is neither possible nor optimal to construct models at the resolution of the hard data. Models are generated at some intermediate geologic modeling scale, and then scaled to an even coarser resolution for process performance. A common goal of geostatistics is the creation of detailed numerical three-dimensional geologic models that simultaneously account for a wide range of relevant data of varying degrees of resolution, quality, and certainty. Much of geostatistics relates to data calibration and reconciling data types at different scales.

At any instance in geologic time, there is a single true distribution of variables over each study area. This true distribution is the result of a complex succession of physical, chemical, and biological processes. Although some of these processes may be understood quite well, we do not completely understand all of the processes and their interactions, and could never have access to the boundary conditions in sufficient detail to provide the unique true distribution of properties. We can only hope to create numerical models that mimic the physically significant features. Uncertainty exists because of our lack of knowledge. Geostatistical techniques allow alternative realizations to be generated. These realizations are often combined in a histogram as a model of uncertainty.

Conventional mapping algorithms were devised to create smooth maps to reveal large-scale geologic trends; they are low-pass filters that remove high-frequency property variations. The goal of such conventional mapping algorithms, including splines and inverse distance estimation, is *not* to show the full variability of the variable being mapped. For many practical problems, however, this variability has a large affect on the predicted response. Geostatistical simulation techniques, conversely, are devised with the goal of introducing the full variability, that is, creating maps or realizations that are neither unique nor smooth. Although the small-scale variability of these realizations may mask large-scale trends, geostatistical simulation is more appropriate for most engineering applications.

There are often insufficient data to provide reliable statistics. For this reason, data from analogous, more densely sampled study areas are used to help infer spatial statistics that are impossible to calculate from the available data. There are general features of certain geologic settings that can be transported to other study areas of similar geologic setting. Although the use of analogous data is often essential in geostatistics, it should be critically evaluated and adapted to fit any hard data from the study area.

A sequential approach is often followed for geostatistical modeling. The overall geometry and major layering or zones are defined first, perhaps deterministically. The rock types are modeled within each major layer or zone. Continuous variables are modeled within homogeneous rock types. Repeating the entire process creates multiple equally probable realizations.

A. Random Variables

The uncertainty about an unsampled value z is modeled through the probability distribution of a random variable (RV) Z. The probability distribution of Z after data conditioning is usually location-dependent; hence the notation $Z(\mathbf{u})$, with $\mathbf{u}$ being the coordinate location vector. A random function (RF) is a set of RVs defined over some field of interest, e.g., $Z(\mathbf{u})$, $\mathbf{u} \in$ study area A. Geostatistics is concerned with inference of statistics related to a random function (RF).

Inference of any statistic requires some repetitive sampling. For example, repetitive sampling of the variable $z(\mathbf{u})$ is needed to evaluate the cumulative distribution function: $F(\mathbf{u}; z) = \text{Prob}\{Z(\mathbf{u}) \leq z\}$ from experimental proportions. However, at most, one sample is available at any single location $\mathbf{u}$; therefore, the paradigm underlying statistical inference processes is to trade the unavailable replication at location $\mathbf{u}$ for replication over the sampling distribution of z samples collected at other locations within the same field.

This trade of replication corresponds to the decision of stationarity. Stationarity is a property of the RF model, not of the underlying physical spatial distribution. Thus, it cannot be checked from data. The decision to pool data into statistics across rock types is not refutable *a priori* from data; however, it can be shown inappropriate *a posteriori* if differentiation per rock type is critical to the undergoing study.

II. QUANTIFICATION OF SPATIAL VARIABILITY

A. Declustering

Data are rarely collected with the goal of statistical representivity. Wells are often drilled in areas with a greater

probability of good reservoir quality. Core measurements are taken preferentially from good-quality reservoir rock. These data-collection practices should not be changed; they lead to the best economics and the greatest number of data in portions of the reservoir that contribute the greatest flow. There is a need, however, to adjust the histograms and summary statistics to be representative of the entire volume of interest.

Most contouring or mapping algorithms automatically correct this preferential clustering. Closely spaced data inform fewer grid nodes and, hence, receive lesser weight. Widely spaced data inform more grid nodes and, hence, receive greater weight. Geostatistical mapping algorithms depend on a global distribution that must be equally representative of the entire area being studied.

Declustering techniques assign each datum a weight, $w_i, i = 1, \ldots, n$, based on its closeness to surrounding data. Then the histogram and summary statistics are calculated with the declustering weights. The weights $w_i, i = 1, \ldots, n$, are between 0 and 1 and add up to 1.0. The height of each histogram bar is proportional to the cumulative weight in the interval, and summary statistics such as the mean and variance are calculated as weighted averages. The simplest approach to declustering is to base the weights on the volume of influence of each sample. Determining a global representative histogram is the first step of a geostatistical study. The next step is to quantify the spatial correlation structure.

B. Measures of Spatial Dependence

The covariance, correlation, and variogram are related measures of spatial correlation. The decision of stationarity allows inference of the stationary covariance (also called auto covariance):

$$C(\mathbf{h}) = E[Z(\mathbf{u}+\mathbf{h}) \cdot Z(\mathbf{u})] - m^2,$$

where m is the stationary mean. This is estimated from all pairs of z-data values approximately separated by vector $\mathbf{h}$. At $\mathbf{h} = 0$ the stationary covariance $C(0)$ equals the stationary variance σ^2. The standardized stationary correlogram (also called auto correlation) is defined as

$$\rho(\mathbf{h}) = C(\mathbf{h})/\sigma^2.$$

Geostatisticians have preferred another two-point measure of spatial correlation called the variogram:

$$2\gamma(\mathbf{h}) = E\{[Z(\mathbf{u}+\mathbf{h}) - Z(\mathbf{u})]^2\}$$

The variogram does not call for the mean m or the variance σ^2; however, under the decision of stationarity the covariance, correlogram, and variogram are equivalent tools for characterizing two-point correlation:

$$C(\mathbf{h}) = \sigma^2 \cdot \rho(\mathbf{h}) = \sigma^2 - \gamma(\mathbf{h})$$

This relation depends on the model decision that the mean and variance are constant and independent of location. These relations are the foundation of variogram interpretation. That is, (1) the "sill" of the variogram is the variance, which is the variogram value that corresponds to zero correlation; (2) the correlation between $Z(\mathbf{u})$ and $Z(\mathbf{u}+\mathbf{h})$ is positive when the variogram value is less than the still; and (3) the correlation between $Z(\mathbf{u})$ and $Z(\mathbf{u}+\mathbf{h})$ is negative when the variogram exceeds the sill.

C. Anisotropy

Spatial continuity depends on direction. Anisotropy in geostatistical calculations is *geometric*, that is, defined by a triaxial Cartesian system of coordinates. Three angles define orthogonal x, y, and z coordinates and then the components of the distance vectors are scaled by three range parameters to determine the scalar distance, that is,

$$h = \sqrt{\left(\frac{h_x}{a_x}\right)^2 + \left(\frac{h_y}{a_y}\right)^2 + \left(\frac{h_z}{a_z}\right)^2},$$

where h_x, h_y, and h_z are the components of a vector $\mathbf{h}$ in three-dimensional coordinate space and a_x, a_y, and a_z are scaling parameters in the principal directions. Contour lines of equal "distance" follow ellipsoids. The use of z for the random variable and a coordinate axis is made clear by context. The three x, y, and z coordinates must be aligned with the principal directions of continuity. A coordinate rotation may be required.

The directions of continuity are often known through geologic understanding. In case of ambiguity, the variogram may be calculated in a number of directions. A *variogram map* could be created by calculating the variogram for a large number of directions and distances; then, the variogram values are posted on a map where the center of the map is the lag distance of zero.

D. Variogram Modeling

The variogram is calculated and displayed in the principal directions. These experimental directional variogram points are not used directly in subsequent geostatistical steps such as kriging and simulation; a parametric variogram model is fitted to the experimental points. There are two reasons why experimental variograms must be modeled: (1) there is a need to interpolate the variogram function for $\mathbf{h}$ values where too few or no experimental data pairs are available, and (2) the variogram measure $\gamma(\mathbf{h})$ must have the mathematical property of "positive definiteness" for the corresponding covariance model—that

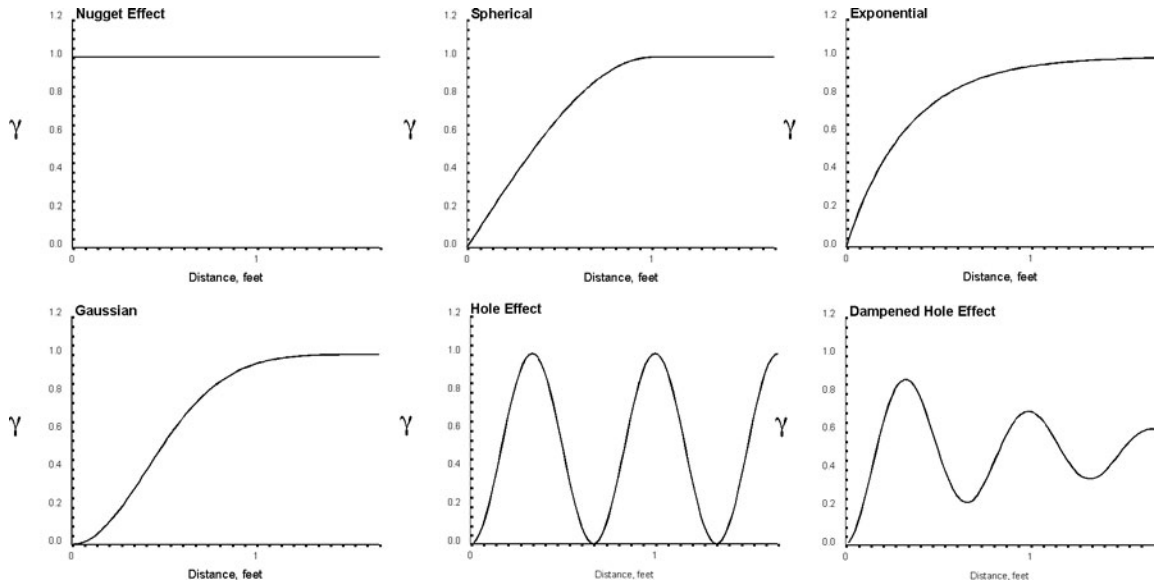


FIGURE 1 Typical variogram structures that are combined together in nested structures to fit experimental variograms. Anisotropy, that is, different directional variograms, are brought to the same distance units by geometric scaling.

is, we must be able to use the variogram and its covariance counterpart in kriging and stochastic simulation. For these reasons, geostatisticians have fitted sample variograms with specific known positive definite functions such as the spherical, exponential, Gaussian, and hole-effect variogram models (see Fig. 1).

A variogram model can be constructed as a sum of known positive-definite licit variogram functions called nested structures. Each nested structure explains a fraction of the variability. All nested structures together describe the total variability, σ^2. Interactive software is typically used to fit a variogram model to experimental points in different directions.

III. SPATIAL REGRESSION OR KRIGING

A. Point Estimation

An important application of geostatistics is to calculate estimates at unsampled locations. The basic idea is to propose a liner estimate of the residual from the mean:

$$z^*(\mathbf{u}) - m(\mathbf{u}) = \sum_{\alpha=1}^{n} \lambda_\alpha \cdot [z(\mathbf{u}_\alpha) - m(\mathbf{u}_\alpha)],$$

where $z^*(\mathbf{u})$ is an estimate made with n data, $m(\mathbf{u})$ is the mean value known at all locations, and λ_α, $\alpha = 1, \ldots, n$, are weights that account for how close the n data are to the location being estimated and how redundant the data are with each other. The weights could be assigned inversely proportional to the distance between the data $\mathbf{u}_\alpha$ and location being estimated, $\mathbf{u}$; however, a better procedure is to use the variogram and minimize the error variance.

B. Simple Kriging

Least-squares optimization has been used for many years. The idea, proposed by early workers in geostatistics, was to calculate the weights to be optimum in a minimum squared error sense, that is, minimize the squared difference between the true value $z(\mathbf{u})$ and the estimator $z^*(\mathbf{u})$. Of course, the true values are known only at the data locations, *not* at the locations being estimated. Therefore, as is classical in statistics, the squared error is minimized in the expected value.

The geostatistical technique known as simple kriging is a least-squares regression procedure to calculate the weights that minimize the squared error. A set of n equations must be solved to calculate the n weights:

$$\sum_{\beta=1}^{n} \lambda_\beta C(\mathbf{u}_\beta - \mathbf{u}_\alpha) = C(\mathbf{u} - \mathbf{u}_\alpha), \quad \alpha = 1, \ldots, n.$$

Recall that $C(\mathbf{h}) = \sigma^2 - \gamma(\mathbf{h})$; therefore, knowledge of the variogram model permits calculation of all needed covariance terms. The left-hand side contains all of the information related to redundancy in the data, and the right-hand side contains all of the information related to closeness of the data to the location being estimated. Kriging is the best estimator in terms of minimum error variance.

Kriging is an exact estimator; that is, the kriging estimator at a data location will be the data value. The minimized error variance or *kriging variance* can be calculated for all estimated locations:

$$\sigma_K^2(\mathbf{u}) = \sigma^2 - \sum_{\alpha=1} \lambda_\alpha \cdot C(\mathbf{u} - \mathbf{u}_\alpha),$$

where the kriging variance is the global variance, σ^2, in the presence of no local data and 0 at a data location. The kriging estimates and kriging variance can be calculated at each location and posted on maps.

C. Constrained Kriging

The basic estimator written in Section III.A requires the mean $m(\mathbf{u})$ at all locations. A number of techniques have been developed in geostatistics to relax this requirement. Ordinary kriging, for example, assumes that the mean m is constant and unknown. A constraint is added to the kriging equations to enforce the sum of the weights to equal 1, which amounts to estimating the mean at each location. Universal kriging assumes the mean follows a particular parametric shape; the parameters are estimated at each location. These constrained versions of kriging make a different decision regarding stationarity.

D. Multiple Variables

The term *kriging* is traditionally reserved for linear regression using data with the same variable as that being estimated. The term *cokriging* is reserved for linear regression that also uses data defined on different attributes. For example, the porosity value $z(\mathbf{u})$ may be estimated from a combination of porosity samples and related acoustic impedance values, $y(\mathbf{u})$. Kriging requires a model for the Z variogram. Cokriging requires a *joint* model for the matrix of variogram functions including the Z variogram, $\gamma_Z(\mathbf{h})$, the Y variogram, $\gamma_Y(\mathbf{h})$, and the cross Z–Y variogram $\gamma_{Z-Y}(\mathbf{h})$. When K different variables are considered, the covariance matrix requires K^2 covariance functions. The inference becomes demanding in terms of data and the subsequent joint variogram modeling; however, cokriging provides the minimum error-variance estimator of the variable at an unsampled location using multiple data variables.

E. Smoothing

Kriging estimates are smooth. The kriging variance is a quantitative measure of the smoothness of the kriging estimates. There is no smoothing when kriging at a data location, $\sigma_K^2 = 0$. There is complete smoothness when kriging with data far from the location being estimated; the estimate is equal to the mean and the kriging variance is the full variance, $\sigma_K^2 = \sigma^2$. This nonuniform smoothing of kriging is the largest shortcoming of kriging for map making. A map of kriging estimates gives an incorrect picture of variability, and calculated results such as recoverable reserves and flow properties are wrong. Simulation corrects for the smoothing of kriging.

IV. SIMULATION

A. Sequential Gaussian Simulation

The idea of simulation is to draw multiple, equally probable realizations from the random function model. These realizatios provide a *joint* measure of uncertainty. Each realization should reproduce (1) the local data at the correct scale and measured precision, (2) the global stationary histogram within statistical fluctuation, and (3) the global stationary variogram or covariance within statistical fluctuation. There is much discussion in the geostatistical literature about different random function models. The most commonly used, however, is the multivariate Gaussian model. The data are first transformed so that the global stationary histogram is Gaussian or normal. Then, all multivariate distributions of n points taken at a time are assumed to follow the mathematically congenial Gaussian distribution. There are many techniques to draw simulations from a multivariate Gaussian random function. The sequential approach gained wide popularity in the 1990s because of its simplicity and flexibility. The sequential Gaussian simulation (SGS) algorithm is as follows.

1. Transform the original Z data to a standard normal distribution (all work will be done in "normal" space). There are different techniques for this transformation. The normal score transformation whereby the normal transform y is calculated from the original variable z as $y = G^{-1}[F(z)]$, where $G(\cdot)$ is the standard normal cumulative distribution function (cdf) and $F(\cdot)$ is the cdf of the original data.
2. Go to a location $\mathbf{u}$ (chosen randomly from the set of locations that have not been simulated yet) and perform kriging to obtain a kriged estimate and the corresponding kriging variance.
3. Draw a random residual $R(\mathbf{u})$ that follows a normal distribution with mean of 0.0 and a variance of $\sigma_K^2(\mathbf{u})$. Add the kriging estimate and residual to get a simulated value. The independent residual $R(\mathbf{u})$ is drawn with classical Monte Carlo simulation.
4. The simulated value is added to the data set and used in future kriging and simulation to ensure that the

variogram between all of the simulated values is correct. A key idea of sequential simulation is to add previously simulated values to the data set.

5. Visit all locations in a random order (return to step 2). There is no theoretical requirement for a random order or path; however, practice has shown that a regular path can induce artifacts. When every grid node has been assigned, the data values and simulated values are back-transformed to real units.

Repeating the entire procedure with a different random number seeds creates multiple realizations. The procedure is straightforward; however, there are a number of implementation issues, including (1) a reasonable three-dimensional model for the mean $m(\mathbf{u})$ must be established, (2) the input statistics must be reliable, and (3) reproduction of all input statistics must be validated.

B. Alternatives to Sequential Approach

Many algorithms can be devised using the properties of the multi-Gaussian distribution to create stochastic simulations: (1) matrix approaches (LU decomposition), which are not used extensively because of size restrictions (an $N \times N$ matrix must be solved, where N could be in the millions for reservoir applications); (2) turning bands methods, where the variable is simulated on one-dimensional lines and then combined into a three-dimensional model, which is not commonly used because of artifacts; (3) spectral methods using fast Fourier transforms can be CPU-fast, but the grid size N must be a power of 2 and honoring conditioning data requires an expensive kriging step; (4) fractals, which are not used extensively because of the restrictive assumption of self-similarity, and (5) moving-average methods, which are used infrequently due to CPU requirements.

C. Indicator Simulation

The aim of the indicator formalism for categorical variables is to simulate the distribution of a categorical variable such as rock type, soil type, or facies. A sequential simulation procedure is followed, but the distribution at each step consists of estimated probabilities for each category: $p^*(k)$, $k = 1, \ldots, K$, where K is the number of categories. The probability values are estimated by first coding the data as indicator or probability values—that is, an indicator is 1 if the category is present, and 0 otherwise. The Monte Carlo simulation at each step is a discrete category. Requirements for indicator simulation include K variograms of the indicator transforms and K global proportions.

V. SPECIAL TOPICS

A. Object-Based Modeling

Object-based models are becoming popular for creating facies models in petroleum reservoirs. The three key issues to be addressed in setting up an object-based model are (1) the geologic shapes, (2) an algorithm for object placement, and (3) relevant data to constrain the resulting realizations. There is no inherent limitation to the shapes that can be modeled with object-based techniques. Equations, a raster template, or a combination of the two can specify the shapes. The geologic shapes can be modeled hierarchically—that is, one object shape can be used at large scale and then different shapes can be used for internal small-scale geologic shapes. It should be noted that object-based modeling has nothing to do with object-oriented programming in a computer sense.

The typical application of object-based modeling is the placement of abandoned sand-filled fluvial channels within a matrix of floodplain shales and fine-grained sediments. The sinuous channel shapes are modeled by a one-dimensional centerline and a variable cross section along the centerline. Levee and crevasse objects can be attached to the channels. Shale plugs, cemented concretions, shale clasts, and other non-net facies can be positioned within the channels. Clustering of the channels into channel complexes or belts can be handled by large-scale objects or as part of the object-placement algorithm.

Object-based facies modeling is applicable to many different depositional settings. The main limitation is coming up with a suitable parameterization for the geologic objects. Deltaic or deep-water lobes are one object that could be defined. Eolean sand dunes, remnant shales, and different carbonate facies could also be used.

B. Indicator Methods

The indicator approach to categorical variable simulation was mentioned earlier. The idea of indicators has also been applied to continuous variables. The key idea behind the indicator formalism is to code all of the data in a common format, that is, as *probability* values. The two main advantages of this approach are (1) simplified data integration because of the common probability coding, and (2) greater flexibility to account for different continuity of extreme values. The indicator approach for continuous data variables requires significant additional effort versus Gaussian techniques.

The aim of the indicator formalism for continuous variables is to estimate directly the distribution of uncertainty $F^*(z)$ at unsampled location $\mathbf{u}$. The cumulative distribution function is estimated at a series of threshold values:

z_k, $k = 1, \ldots, K$. The indicator coding at location $\mathbf{u}_\alpha$ for a particular threshold z_k is

$$i(\mathbf{u}_\alpha; z_k) = \text{Prob}[Z(\mathbf{u}_\alpha) \leq z_k] = \begin{cases} 1, & \text{if } z(\mathbf{u}_\alpha) \leq z_k, \\ 0, & \text{otherwise.} \end{cases}$$

All hard data $z(\mathbf{u}_\alpha)$ are coded as discrete zeros and ones. Soft data can take values between zero and one. The indicator transform for a threshold less than the data value is zero, since there is no probability that the data value is less than the threshold; the indicator transform for a very high threshold is one, since the data value is certainly less than the threshold.

The cumulative distribution function at an unsampled location at threshold z_k can be estimated by kriging. This "indicator kriging" or IK requires a variogram measure of correlation corresponding to each threshold z_k, $k = 1, \ldots, K$. The IK process is repeated for all K threshold values that discretize the interval of variability of the continuous attribute Z. The distribution of uncertainty, built from assembling the K indicator kriging estimates, can be used for uncertainty assessment or simulation.

C. Simulated Annealing

The method of simulated annealing is an optimization technique that has attracted significant attention. The task of creating a three-dimensional numerical model that reproduces some data is posed as an optimization problem. An objective function measures the mismatch between the data and the numerical model. An initial random model is successively perturbed until the objective function is lowered to zero. The essential contribution of simulated annealing is a prescription for when to accept or reject a given perturbation. This acceptance probability distribution is taken from an analogy with the physical process of annealing, where a material is heated and then slowly cooled to obtain low energy.

Simulated annealing is a powerful optimization algorithm that can be used for numerical modeling; however, it is more difficult to apply than kriging-based methods because of difficulties in setting up the objective function and choosing many interrelated parameters such as the annealing schedule. Therefore, the place of simulated annealing is not for conventional problems where kriging-based simulation is adequate. Simulated annealing is applicable to difficult problems that involve (1) dynamic data, (2) large-scale soft data, (3) multiple-point statistics, (4) object placement, or (5) special continuity of extremes.

D. Change of Support

Reconciling data from different scales is a long-standing problem in geostatistics. Data from different sources, including remotely sensed data, must all be accounted for in the construction of a geostatistical reservoir model. These data are at vastly different scales, and it is wrong to ignore the scale difference when constructing a geostatistical model. Geostatistical scaling laws were devised in the 1960s and 1970s primarily in the mining industry, where the concern was mineral grades of selective mining unit (SMU) blocks of different sizes. These techniques can be extended to address problems in other areas, subject to implicit assumptions of stationarity and linear averaging.

The first important notion in volume-variance relations is the spatial or dispersion variance. The dispersion variance $D^2(a, b)$ is the variance of values of volume a in a larger volume b. In a geostatistical context, all variances are dispersion variances. A critical relationships in geostatistics is the link between the dispersion variance and the average variogram value:

$$D^2(a, b) = \bar{\gamma}(b, b) - \bar{\gamma}(a, a).$$

This tells us how the variability of a variable changes with the volume scale and variogram. The variability of a variable with high short-scale variability decreases quickly, since high and low values average out.

VI. APPLICATIONS AND EXAMPLES

A. Environmental

Figure 2 illustrates some of the geostatistical operations applied to characterize the spatial distribution of lead contamination over a 12,500-ft^2 area. There are five parts to Fig. 2: (1) the upper left shows the location map of the 180 samples—there is no evident clustering that would require declustering; (2) the equal-weighted histogram, at the upper right, shows the basic statistics related to the measurements—note the logarithmic scale; (3) the variogram, shown below the histogram, is of the normal scores transform of the lead data—about 40% of the variability is at very short distances and the remaining 60% of the variability is explained over 4500 ft—the black dots are the experimentally calculated points and the solid line is the fitted model; (4) a map of kriging estimates on a 100-ft^2 grid is shown at the lower left—note the smoothness of the kriging estimates; and (5) a sequential Gaussian simulation (SGS) realization is shown at the lower right—this realization reproduces the 180 sample data, the input histogram, and the variogram model. A set of realizations could be used to assess the probability that each location exceeds some critical threshold of lead concentration.

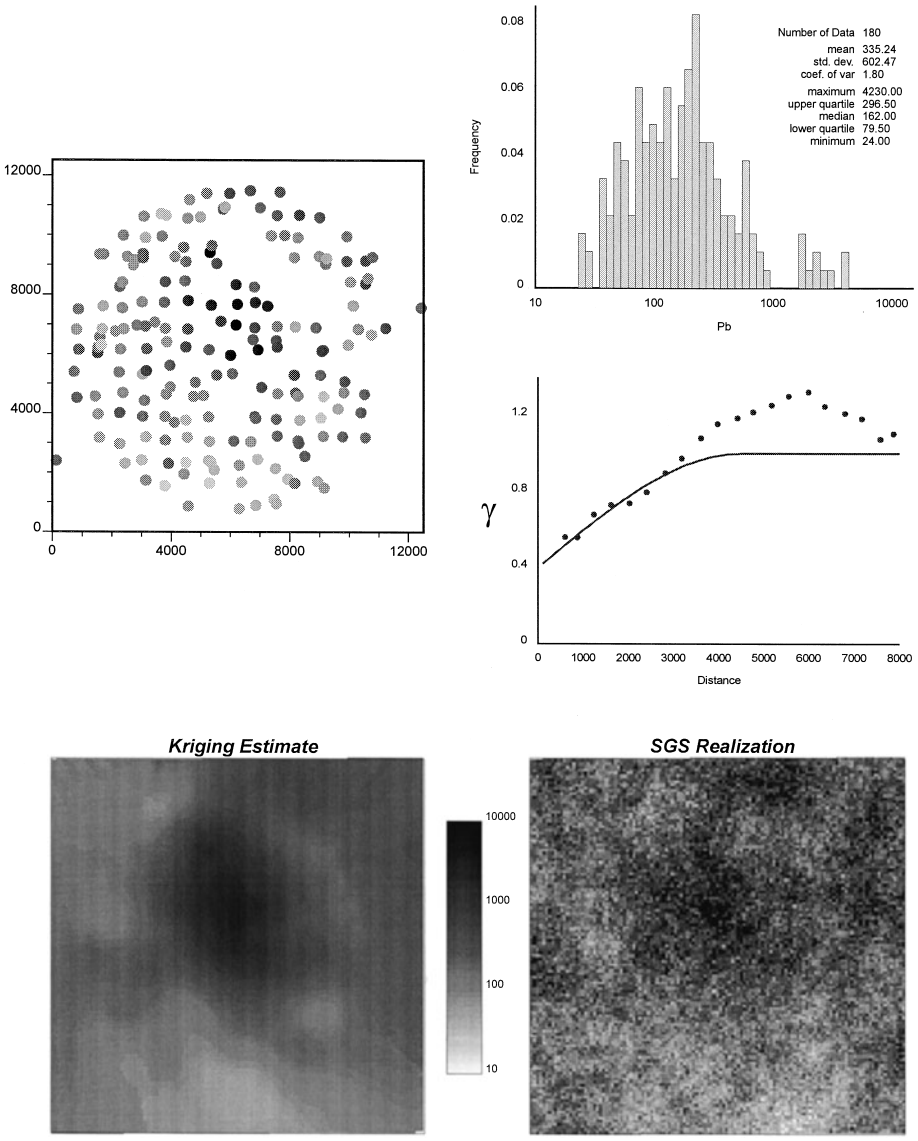


FIGURE 2 Location map (distance units in feet) of 180 samples, histogram of lead concentration, variogram of the normal scores transform (hence the *sill* value of 1.0), a map of kriging estimates on a 100-ft^2 grid and an SGS realization over the same domain.

B. Mining

Figure 3 illustrates an example application to a vein-type mineral deposit. The cross-sectional view at the upper left is a vertical cross section facing west; the vertical coordinate is meters below the surface. The drillhole intersections are clustered in the thickest part of the vein. The polygonal areas of influence plotted on the location map are used for declustering weights. The histogram at the upper right of the figure considers the declustering weights. The variogram is shown below the histogram. Two nested structures were used to fit this variogram. One sequential Gaussian realization is shown at the lower left; 150 realizations were generated. The probability of exceeding 1-m thickness is plotted at the lower right. The black locations are where the vein is measured to be greater than 1 m in thickness (probability of 1), and the white locations are where the vein is measured to be less than 1 m (probability of 0).

C. Petroleum

The profile of porosity and permeability from two wells from an offshore petroleum reservoir are shown at the

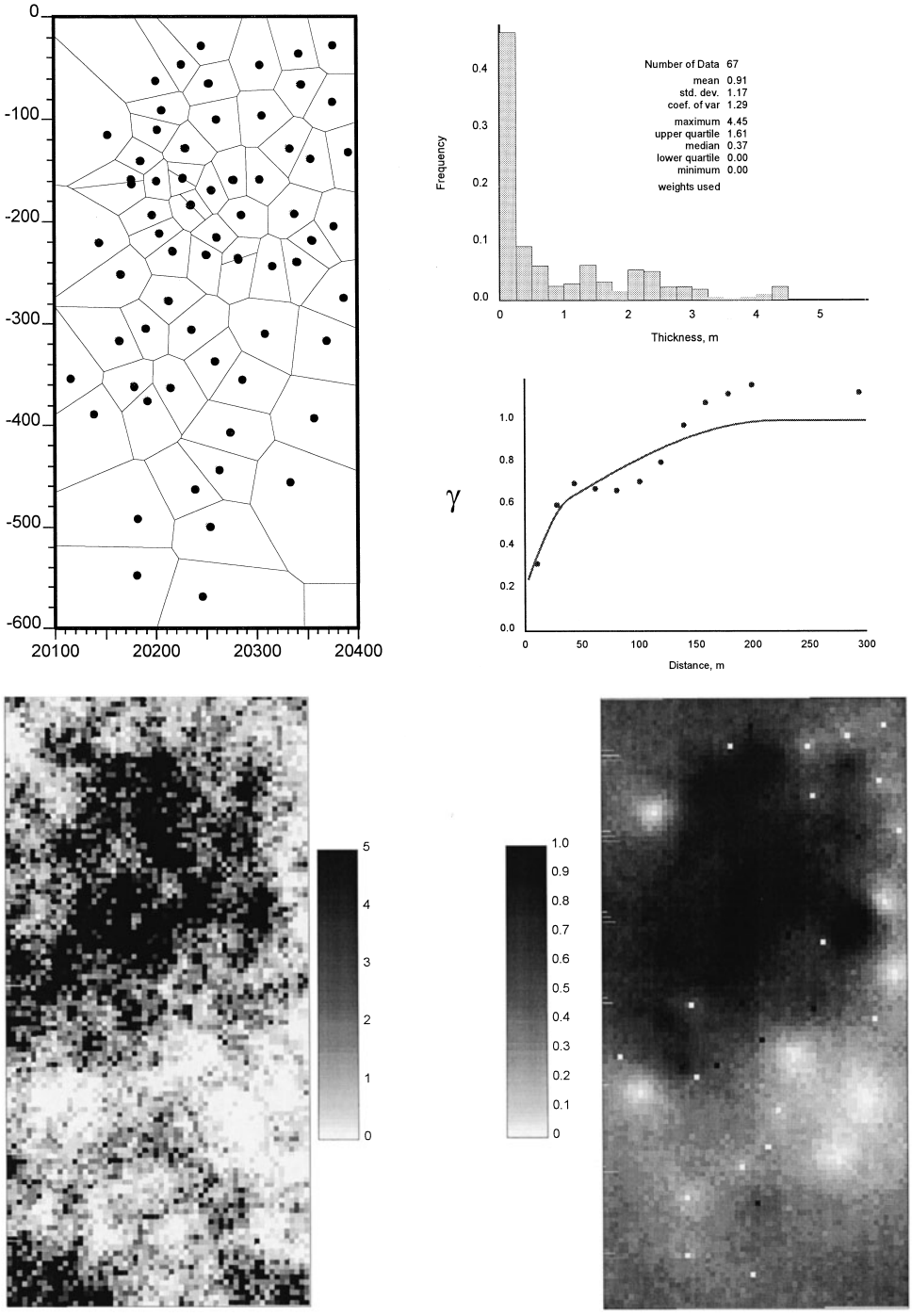


FIGURE 3 Location map (distance units in meters) of 67 drillholes with polygonal areas of influence for declustering weights, histogram of vein thickness, variogram of the normal scores transform, an SGS realization over the same domain, and the probability to exceed 1.0-m thickness calculated from 100 realizations.

bottom of Fig. 4. A porosity and permeability realization are shown at the top. Simulation of porosity and permeability were done simultaneously to reproduce the correlation between these two variables. The vertical variograms were calculated and modeled easily; however, the horizontal variograms are impossible to discern from two wells. A 50:1 horizontal-to-vertical anisotropy was considered from analog data.

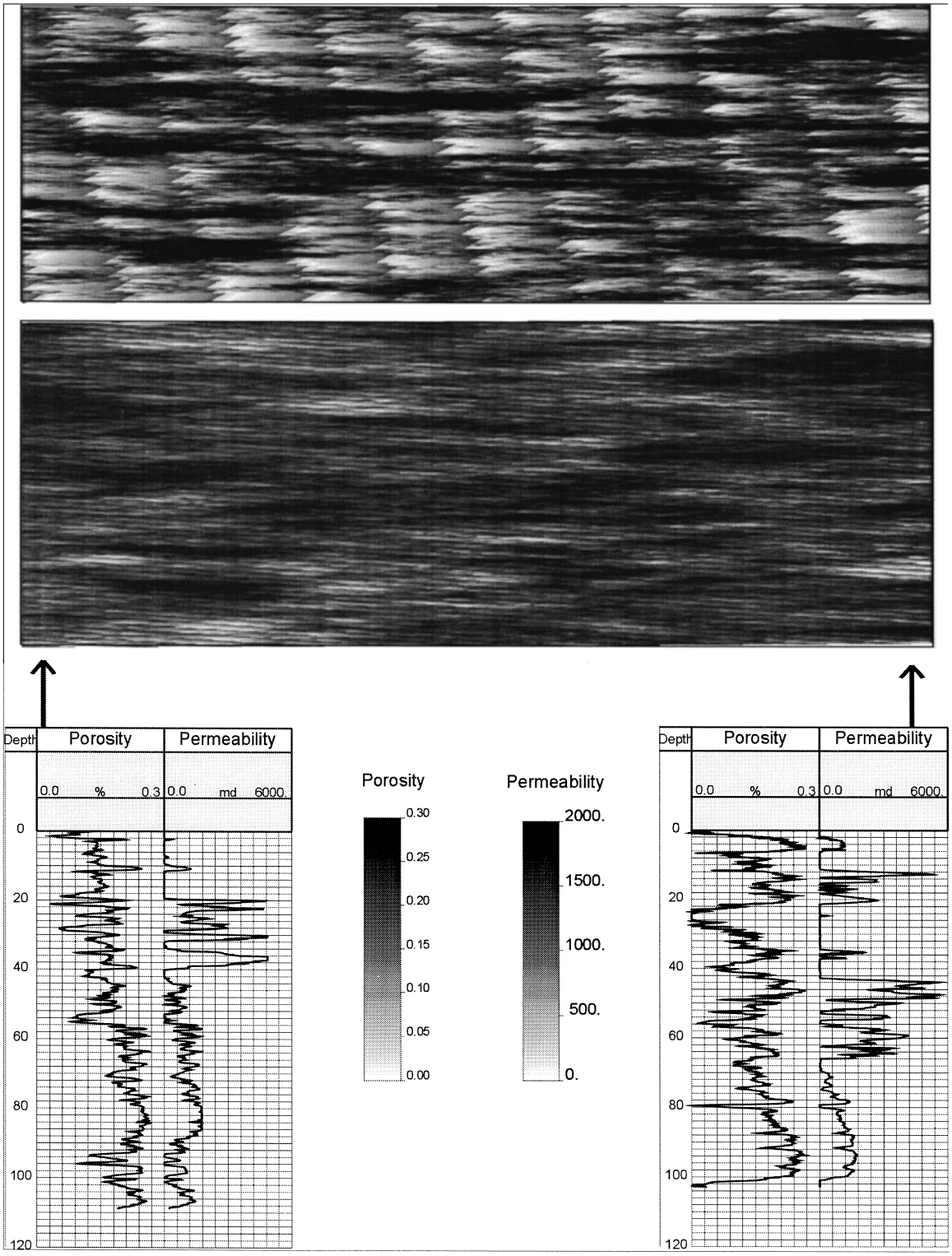


FIGURE 4 Permeability realization (top) and porosity realization (middle) constrained to two wells 600 m apart (shown at the bottom).

SEE ALSO THE FOLLOWING ARTICLES

MINING ENGINEERING • ORE PETROLOGY • PETROLEUM GEOLOGY • STATISTICS, FOUNDATIONS

BIBLIOGRAPHY

Chiles, J. P., and Delfiner, P. (1999). "Geostatistics: Modeling Spatial Uncertainty" (Wiley Series in Probability and Statistics, Applied Probability and Statistics), Wiley, New York.

Cressie, N.(1991). "Statistics for Spatial Data," Wiley, New York.

David, M. (1977). "Geostatistical Ore Reserve Estimation," Elsevier, Amsterdam.

Deutsch, C. V., and Journel, A. G. (1997). "GSLIB: Geostatistical Software Library," 2nd ed., Oxford University Press, New York.

Goovaerts, P. (1997). "Geostatistics for Natural Resources Evaluation," Oxford University Press, New York.

Isaaks, E. H., and Srivastava, R. M. (1989). "An Introduction to Applied Geostatistics," Oxford University Press, New York.

Journel, A. G., and Huijbregts, C. (1978). "Mining Geostatistics," Academic Press, London.

Matheron, G. (1971). "The theory of regionalized variables and its applications," *Cahiers du CMM* **5,** Ecole Nationale Superieure des Mines de Paris, Paris.

Ripley, B. D. (1981). "Spatial Statistics," Wiley, New York.

Global Gravity Modeling

R. Steven Nerem
Colorado Center for Astrodynamics Research

GLOSSARY

Geoid Equipotential surface of the Earth's gravity field that best approximates mean sea level. Over the land, this surface can lie above the surface, or below the surface.

Gravity Force exerted by a mass on another body, typically measured in units of acceleration in milliGals ($1\ \text{gal} = 1\ \text{cm/s}^2$).

Geopotential A representation of the potential of the Earth that satisfies Laplace's equation ($\nabla^2 U = 0$).

Spherical harmonics Means of mathematically representing a variable in terms of trigonometric functions, each having a different amplitude and wavelength, on the surface of a sphere. Analogous to a Fourier representation of a two-dimensional function, but in this case used for representing spherical functions.

GLOBAL GRAVITY MODELING an area of study that attempts to make the best possible estimate of the detailed gravity field of a planet using both terrestrial gravity measurements and more recently, satellite measurements. These models, together with topography models, are used in the fields of geodesy, geophysics, and planetary physics to characterize the internal structure of the planet and the dynamics of planetary interiors. These models also see wide use in the aerospace community for trajectory determination of spacecraft and missiles.

I. DEFINITION OF THE GLOBAL GRAVITY FIELD

The concept of a global gravity field is based on the basic principles of physics, which is at present largely Newtonian mechanics. Newton's Law of Gravitation states that the magnitude of the force between two masses M and m is inversely proportional to the square of the distance (r) between them and may be written as:

$$F = \frac{GMm}{r^2}, \tag{1}$$

where G is the Universal Gravitational Constant ($6.673 \times 10^{-20}\ \text{km}^3/\text{kg s}^2$). In the gravity modeling community, the vector force of gravity is usually represented as the gradient of a potential, where the gradient operator (∇) can be written as:

$$\nabla = \frac{\partial}{\partial x}\hat{i} + \frac{\partial}{\partial y}\hat{j} + \frac{\partial}{z}\hat{k}. \tag{2}$$

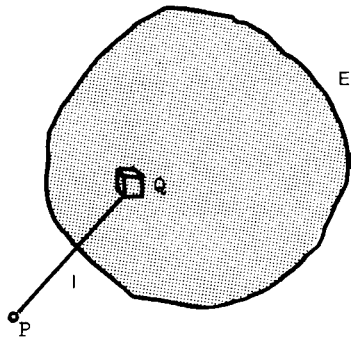


FIGURE 1 Mass distribution M with volume element dm and object at point P (from Fig. 1 of Anderson chapter with $Q = dm$, $E = M$, and $l = \rho$).

The potential is usually represented as an integral over the planet of the contribution of each infintesimal mass dm as (Fig. 1)

$$U = \int_M \frac{Gm}{\rho} dm, \tag{3}$$

where ρ is the distance from the incremental mass dm to a point P outside the Earth. For a spherical homogenous planet (point mass), this reduces to

$$U = \frac{GMm}{r}, \tag{4}$$

where r is the distance from the coordinate system origin to the satellite. The force on the satellite can be computed from the gradient of the potential

$$\vec{F} = \nabla U = \frac{GMm}{r^2}\frac{\vec{r}}{r}, \tag{5}$$

which is the vector equivalent of Eq. (1). This representation is sufficient for points that are a long distance from the planet or when accuracy requirements are not demanding. For more stringent applications, the planet should be considered nonhomogenous. Then, $1/\rho$ in Eq. (3) is expanded in terms of spherical harmonic functions, and the integral results in a constant coefficient associated with each harmonic in the expansion:

$$U = \frac{GM}{r}\left[1 + \sum_{l=2}^{\infty}\sum_{m=0}^{l}\left(\frac{a_e}{r}\right)^l \bar{P}_{lm}(\sin\phi) \times (\bar{C}_{lm}\cos m\lambda + \bar{S}_{lm}\sin m\lambda)\right], \tag{6}$$

where a_e is the equatorial radius of the planet, ϕ and λ are the latitude and longitude of the satellite, l and m are the degree and order of the spherical harmonic expansion, $P_{lm}(\sin\phi)$ are the fully normalized Legendre associated functions of degree l and order m (for $m = 0$, these are the Legendre polynomials of degree l), and the C_{lm}/S_{lm} are the fully normalized spherical harmonic coefficients describing the spatial variations of the Earth's potential field. The constants GM, a_e, and C_{lm}/S_{lm} will of course be unique for each planet, and are referred to as a gravity models for that planet. For Earth, $GM = 398600.4415$ km^3/s^2 and $a_e = 6378.1363$ km, and the first few values of the spherical harmonic coefficients are shown in Table I. For a given spherical harmonic degree l, the corresponding half-wavelength spatial resolution on the Earth's surface is approximately given by $20000/l$ km. The degree l coefficients of the spherical harmonic potential (6) are usually assumed to be zero if the mass center is assumed to coincide with the origin of the coordinate system being used. While the expansion is theoretically infinite, in practice it is complete to $l = 360$ or less, depending on the spatial resolution of the data used to determine the spherical harmonic coefficients. Higher-resolution global gravity models are usually presented in gridded form rather than as a spherical harmonic representation.

The gravity field of the Earth also varies as a function of time as it deforms due to the gravitational effects of the Sun and the Moon, resulting in solid Earth and ocean tides. The tidal effects have been reasonably well determined, mainly because they occur at well-known astronomical frequencies. In addition, the gravity field varies slightly as mass is redistributed on its surface and in its interior. The phenomena of postglacial rebound refers to the slow rebound of the Earth's crust, predominantly in North America and Scandinavia, due to the melting of the ice sheets at the end of the last ice age 10,000 years ago. This rebound causes small secular changes in the spherical harmonic coefficients of the gravity field. In addition, the gravity field varies as water mass moves amongst the continents, oceans, and atmosphere [Wahr *et al.*, 1998]. There are a host of smaller effects.

TABLE I Current Global Geopotential Models

Model	Date	NMAX	Data used[a]
GEM9	1977	20	S
Rapp	1978	180	$S + G(A) + G(T)$
SAO	1980	30	$S + A + G(T)$
GEM 10B	1981	36	$S + A + G(T)$
GEM 10C	1981	180	$S + A + G(T)$
Rapp	1981	180	$S + G(A) + G(T)$
GEML2	1982	20	S
GRIM3B	1983	36	$S + G(A) + G(T)$
GRIM3-L1	1984	36	$S + G(A) + G(T)$

[a] S is satellite orbit data, A is satellite altimetry data, $G(A)$ is gravity anomaly data derived from satellite altimetry, and $G(T)$ is terrestrially measured gravity anomalies (surface gravimeter-based data).

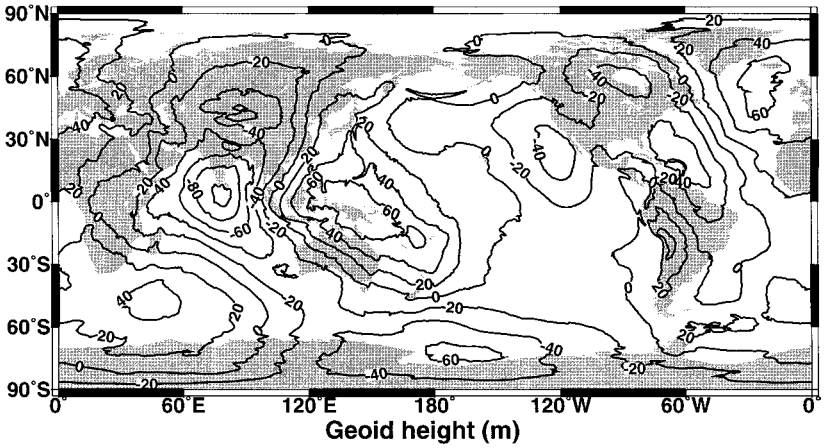


FIGURE 2 Contour map of the geoid height complete to degree and order 360 from the EGM-96 gravity model (From Lemoine, F. G. *et al.* (1998). "The Development of the Joint NASA GSFG and the National Imagery and Mapping Agency (NIMA) Geopotential Model EGM96, NASA Goddard Space Flight Center, Greenbelt, MD.)

II. METHOD OF MEASUREMENT

Until the launch of Sputnik, the only method available to measure the Earth's gravity field was through the use of surface gravimeters, and the first crude representations of the global gravity field were assembled by merging together surface gravimetric measurements from around the world. Almost immediately after entering the satellite era, the long wavelength components of the global gravity field were determined by measuring the gravitational perturbations to satellite orbits using ground-based tracking data. This is how the slight "pear shape" of the Earth was first determined (the "oblateness" of the Earth was previously known). In the late 1970s, satellite altimeter measurements were used to study the Earth's gravity field over the oceans, since the ocean surface largely conforms (to within ± 1 m) to the geoid. At present, the most comprehensive models of the global gravity field are determined from a combination of satellite tracking data (collected from dozens of different satellites since the beginning of the satellite era), satellite altimeter data, and surface gravity data (Nerem *et al.*, 1995).

Maps of the global gravity field are usually represented either in the form of the geoid or as gravity anomalies. The geoid is defined as the height of the equipotential surface of the Earth's gravity field with most closely corresponds to mean sea level. The acceleration of gravity on this surface is everywhere the same. Because the ocean is a fluid, it adjusts itself to conform to the geoid, with the exception of the ± 1 m deviations caused by the ocean currents. The geoid height can be found by chosing an appropriate constant value for the potential, U_0, and then determining the radius r from the expression for the geopotential in Eq. (5) plus the rotational potential ($\frac{1}{2}\|\vec{\omega} \times \vec{r}\|$). It is normally expressed relative to the height of a reference ellipsoid which best fits the shape of the Earth. As shown

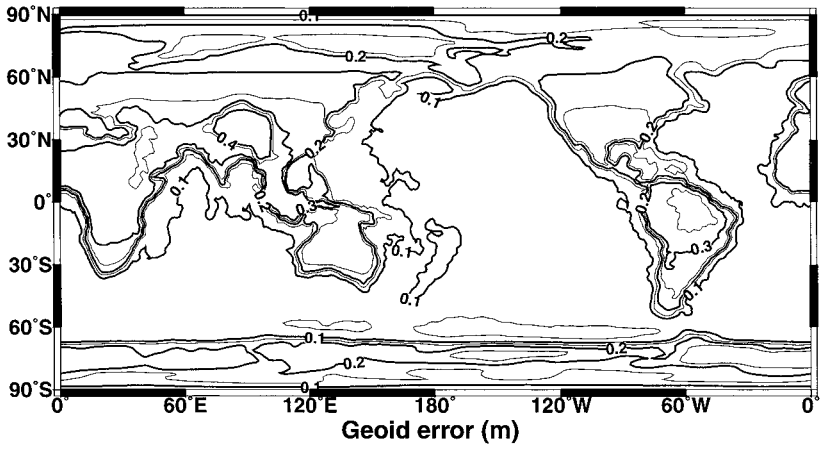


FIGURE 3 Estimates of errors in the geoid height for the EGM-96 model.

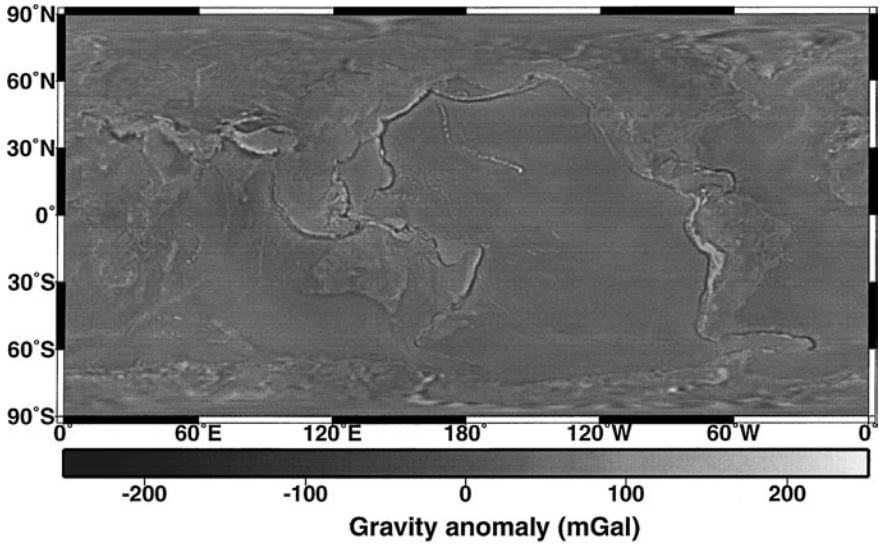


FIGURE 4 Gravity anomalies computed from the EGM-96 gravity model complete to degree and order 360.

in Fig. 2, the geoid height varies by ±100 m relative to the reference ellipsoid, and at long wavelengths mainly reflects density anomalies deep within the Earth. A map of the error in our current knowledge of the geoid is shown in Fig. 3. The errors are lowest over the oceans where we have satellite altimeter measurements, and highest over land areas where surface gravity observations are not available or suffer from poor accuracy. Gravity anomalies are the total gravitational acceleration at a given location minus the acceleration described by the reference ellipsoid, which varies only with latitude. Gravity anomalies are generally generally expressed in milliGals, where 1 Gal = 1 cm/s^2, as shown in Fig. 4. Gravity anomalies, which are "rougher" than the geoid, are better for representing

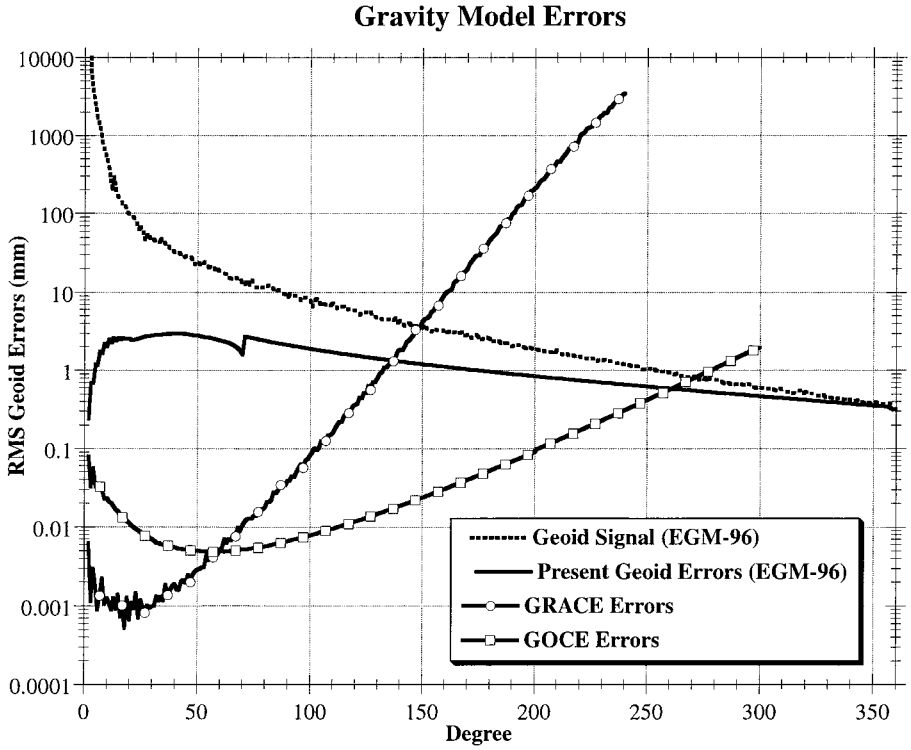


FIGURE 5 Estimates of the errors in our current knowledge of the geoid versus spherical harmonic degree versus the expected errors from two future satellite gravity missions, GRACE and GOCE.

the fine scale density variations near the surface of the Earth.

III. FUTURE DEDICATED SATELLITE GRAVITY MISSIONS

The field of global gravity field determination is entering a new era as satellite missions dedicated to measuring the Earth's gravity field are being developed and flown. In the next few years, the Gravity Recovery and Climate Experiment (GRACE) and the Global Ocean Circulation Explorer (GOCE) will be launched. GRACE will use precise microwave measurements between two satellites flying at an altitude of approximately 450 km to precisely map the Earth's gravity field. In addition, GRACE will be able to detect temporal variations of the Earth's gravity field, which after tidal variations are removed, are predominantly due to water mass being redistributed on the surface of the Earth (snow, ice, ground water, aquifers, etc.), in the atmosphere (water vapor), and in the ocean. It is expected that GRACE will be capable of making monthly estimates of the gravity field with a spatial resolution of ~300–500 km and an accuracy of 1 cm equivalent water thickness (Dickey *et al.*, 1997). GOCE will consist of a single satellite carrying a gravity gradiometer, which will directly measure the gravity gradient (spatial derivative of gravity) in three axes. While GOCE will likely not have enough sensitivity to detect temporal gravity variations at long wavelengths, it will provide a much better determination of the static gravity field that can be provided by GRACE alone. The primary objective for the GOCE mission is to improve our knowledge of the geoid over the oceans to allow detailed studies of ocean circulation using satellite altimetry measurements. The expected errors in our knowledge of the Earth's gravity derived from each of these future missions is shown in Fig. 5.

IV. APPLICATIONS

Global gravity field models are used in a wide variety of applications in geophysics, oceanography, geodesy, and engineering (Fig. 6). Geodesists interested in measuring

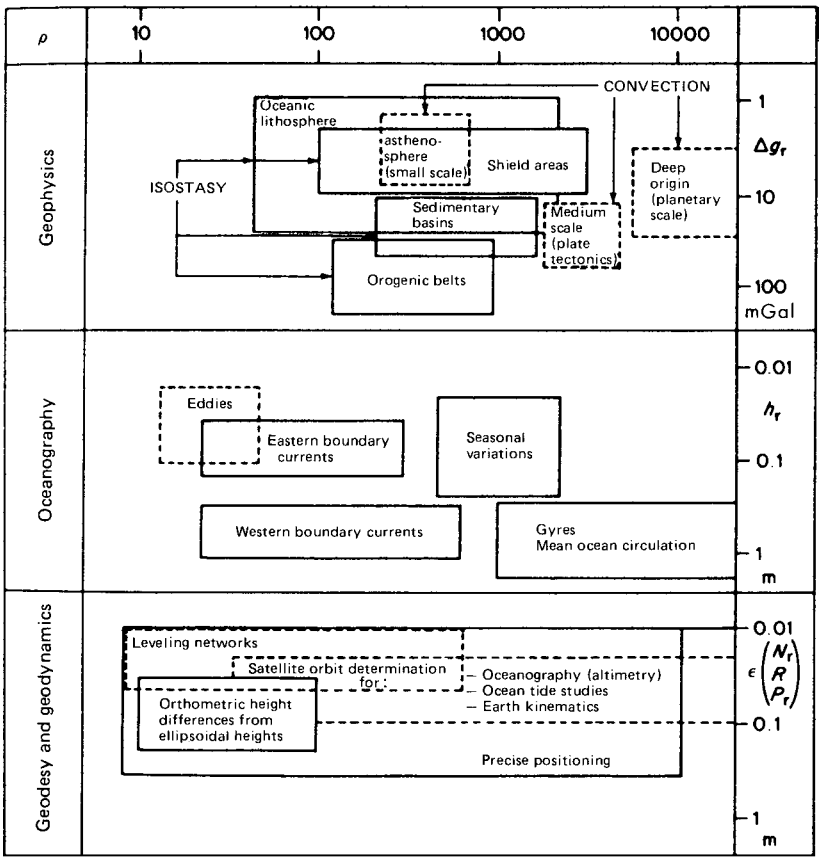


FIGURE 6 Magnitude of various quantities related to Earth global gravity anomalies at different wavelengths. Δg_r is the gravity anomaly in mGals, h_t is the relative variation of ocean surface height in meters, $\varepsilon(N_t)$ is the accuracy of the geoid height in meters, $\varepsilon(R)$ is the accuracy of the radial component of a typical artificial satellite orbit in meters, and $\varepsilon(P_t)$ is the relative position accuracy typically obtained using geodetic techniques.

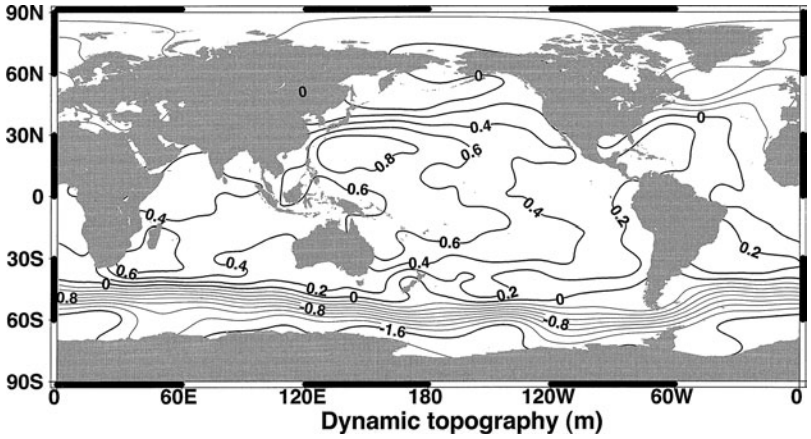


FIGURE 7 Map of the ocean dynamic topography determined using TOPEX/Poseidon satellite altimeter data relative to the EGM-96 geoid model (Lemoine *et al*., 1998). The contour lines are parrallel to the direction of the ocean currents, which move clockwise in the northern hemisphere, and counter-clockwise in the southern hemisphere.

orthometric heights (relative to the geoid or "mean sea level") using the Global Positioning System (GPS) must know the geoid height at the desired location, since GPS provides absolute heights. Oceanographers need to know the geoid height in order to measure the ocean circulation using satellite altimetry, since slope of the difference between the height of the ocean surface and the geoid is directly related to the geostrophic velocity of the ocean currents (Fig. 7). Geophysicists use gravity field models to study the internal structure of the Earth and as a geophysical exploration tool. Gravity field models are also fundamental to accurately computing the trajectories of Earth orbiting satellites, which is very important when making geodetic measurement from space (such as satellite altimetry), and computing ballistic missile trajectories.

Measurements of temporal variations of the Earth's gravity field largely represent the redistribution of water mass in the Earth system and a variety of temporal and spatial scales (Fig. 8), and are of interest to solid Earth geophysicist, hydrologists, meteorologists, oceanographers, and glaciologists, among others. At present, temporal gravity variations derived from satellite measurements have only been detected at wavelengths longer than

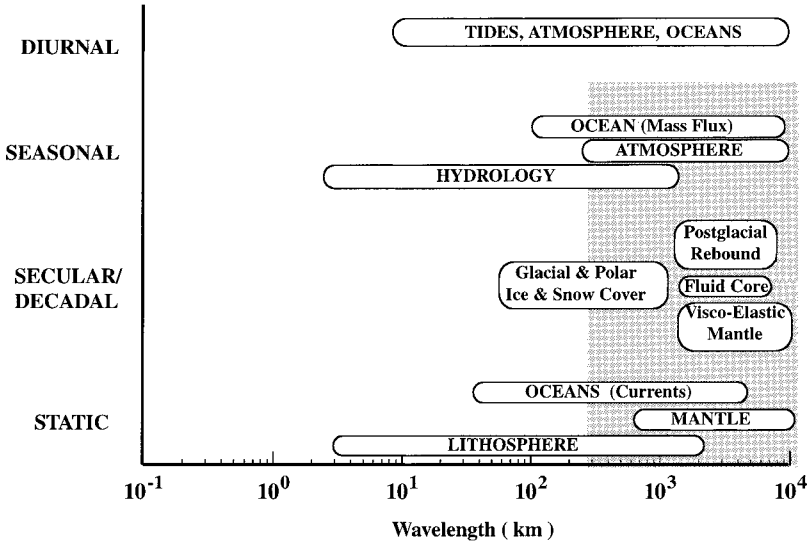


FIGURE 8 Representation of the typical spatial and temporal scales of non-tidal temporal gravity variations on the Earth. The shaded area is the temporal and spatial scales that the GRACE satellite mission is expected to resolve.

10,000 km; however, the GRACE mission should usher in a new paradijm in this field. Thus, the anticipated future improvements to the global gravity model will benefit a wide array of science and engineering applications.

Gravity modeling is also our principal tool for discerning the internal structure of the planets (Nerem *et al.*, 1995). Planetary gravity models are determined from Earth-based tracking of spacecraft orbiting the planet. Combined with topography models, often determined using radar or laser altimetry from these spacecraft, much can be learned about the structure of the planet, such as crustal thickness, density anomalies, and composition of the core. In recent years, precise gravity and topography models have been determined for Venus, the Moon, Mars, and the asteroid Eros using such measurements. Temporals gravity variations are also important for understanding planetary dynamics. The planet Mars is thought to have significant variations in its oblateness due to the annual accumulation/melting of ice at its poles. In addition the tidal variation of gravity and topography on Jupiter's moon Europa is thought to hold the key to determine if a subsurface ocean lies beneath Europa's icy shell.

SEE ALSO THE FOLLOWING ARTICLES

GEODESY • GRAVITATIONAL WAVE DETECTORS • GRAVITATIONAL WAVE PHYSICS • MECHANICS, CLASSICAL • REMOTE SENSING FROM SATELLITES

BIBLIOGRAPHY

Dickey, J. O., Bentley, C. R., Bilham, R., Carton, J. A., Eanes, R. J., Herring, T. A., Kaula, W. M., Lagerloef, G. S. E., Rojstaczer, S., Smith, W. H. F., van den Dool, H. M., Wahr, J. M., and Zuber, M. T. (1997). "Satellite Gravity and the Geosphere: Contributions to the Study of the Solid Earth and Its Fluid Envelope," pp. 112, National Research Council, Washington, DC.

Lemoine, F. G., Kenyon, S. C., Factor, J. K., Trimmer, R. G., Pavlis, N. K., Chinn, D. S., Cox, C. M., Klosko, S. M., Luthcke, S.B., Torrence, M. H., Wang, Y. M., Williamson, R. G., Pavlis, E. C., Rapp, R. H., and Olson T. R. (1998). "The Development of the Joint NASA GSFC and the National Imagery and Mapping Agency (NIMA) Geopotential Model EGM96," NASA Goddard Space Flight Center, Greenbelt, MD.

Nerem, R. S., Jekeli, C., and Kaula, W. M. (1995). "Gravity field determination and characteristics: Retrospective and prospective," *J. Geophys. Res.* **100**(B8), 15053–15074.

Wahr, J., Molenaar, M., and Bryan, F. (1998). "Time variability of the Earth's gravity field: Hydrological and oceanic effects and their possible detection using GRACE," *J. Geophys. Res.* **103**(B12), 30, 205–230.

Microscopy

Richard J. Evans
Becton-Dickinson & Company

GLOSSARY

Anisotropic Classification of materials exhibiting different properties in different directions.
Auger Electron beam instrument capable of chemical analyses on very shallow sample depths (5–20 Å).
Birefringence Optical phenomena of anisotropic substances quantitatively measured as the difference between the refractive indices of the substances.
Crystal Solid three-dimensional material having a repeating three-dimensional array of atoms, ions, or molecules.
Isotropic Classification of materials exhibiting identical physical properties in all directions.
Molecular optical laser examiner (MOLE) Instrument combining a laser excitation source with a microscope, resulting in a raman spectrum which can be used to "fingerprint" chemically small particles or thin films of complex compounds.
Optical crystallography Study of crystals and their effect on light, specifically polarized light.
Polarized light microscope (PLM) Instrumentation used to study optical phenomena as they relate to the chemistry and condition of a material illuminated by polarized light.
Refractive index Measure of the change of the speed of light as it passes from air (vacuum) through a given medium, expressed as a ratio of the two.
Scanning electron microscope (SEM) Electron beam instrument used to observe surface features of materials at great resolution and depth of field, usually used in conjunction with X-ray spectrometers to obtain chemical analytical information from the specimen.
Transmission electron microscope (TEM) Electron beam instrument analogous to light microscopes in that the radiant cone (electrons versus light) is transmitted through the sample. Only very thin samples (<0–10 μm) can be observed in any detail. X-ray spectrometers can also be used with the TEM.
X-ray spectrometer Instrumentation which measures X-ray generation and segregates energies or wavelengths into a spectrum representing the elemental composition of a sample.

CHEMICAL MICROSCOPY may best be defined as any one of many techniques employed to analyze the chemistry, nature, or behavior of materials on a microscopic scale. Classically, chemical microscopy has meant observing optical phenomena of a material or chemical reactions and their end products with a polarized light microscope.

This technique in trained hands yields more information about a material than any other single microanalytical tool. Since the 1970s, however, techniques and instrumentation which can aid the analyst whose material problems may lie beyond the capabilities of the light microscope have been developed; TEM, SEM, ESCA/XPS, and Auger, among others, are the new accessories to which the chemist in search of answers to microscopical problems can also turn.

I. PRIMARY APPLICATIONS

A. Research

There are few types of microscopy that are not within the field of chemistry itself, for example, a biologist observing cell behavior (biochemistry), a geophysicist observing shale structure (physical chemistry), a forensics investigator matching glass particles, or an art conservator validating the authenticity of a painting. All are hoping to learn something about microscopic samples by observing some feature or features of the sample.

During the preliminary research into any new material, whether it is a drug, polymer, metal, or combination of materials, only small amounts of samples are prepared. Often, many different tests must be performed on each of these small samples; therefore, in order for a chemist to obtain chemical information about the sample, he must know or utilize chemical testing on a microscopic scale.

B. Industry

Since the advent of solid-state circuitry and miniaturized electronics, interest in ultramicroanalysis has grown tremendously, primarily with regard to contamination control and contaminant analysis. Exotic new materials which are being developed are often manufactured under cleanroom environments which must be tightly controlled and constantly monitored to ensure the integrity and consistency of the end product. For example, a mere trace of contaminant in a silicon crystal-growing process could jeopardize thousands of circuits.

Similarly, the metallurgist must often analyze small areas or inclusions in an alloy or determine the nature of a fractured surface. These problems can be solved only with microanalytical techniques. In the fields of environmental science and contamination control, the chemist most often analyzes single particles of material a few microns or less in size. Asbestos analysis, for example, as specified by the U.S. Environmental Protection Agency, is to be performed using a polarized light microscope (PLM), which is the principal tool of the chemical microscopist.

Geologists searching for specific types of minerals often determine crystallographic data by PLM to identify minerals. This is a very rapid method to obtain crystallographic data.

C. Forensics

Probably no branch of science is held to higher scrutiny than forensics. In a forensics laboratory, when trace evidence is often all there is available from a crime scene, a trained microscopist is imperative. Drug tests, blood tests, gunshot residue, glass matching, and fiber and hair analysis are all performed on trace evidence using a variety of microchemical techniques, most often with a polarized light microscope.

II. EQUIPMENT AND INSTRUMENTATION

The instrumentation utilized in the field of chemical microscopy can be very simple or highly complex, as can the sample to be analyzed. Therefore, the sample itself frequently determines the type of analysis to be performed; however, no analysis should be performed without first having in mind a question which the analysis will answer or observing the specimen through some type of optical microscope.

A. Optical Systems

1. Stereomicroscope

The simplest optical microscope most frequently encountered in a laboratory is a stereobinocular microscope. This is typically used to magnify an image from 5 to 60 times, and the user sees a three-dimensional, noninverted image. The three-dimensional, or stereo, image is formed by two separate compound microscopes (one for each eye) focused at the same point from two different angles, typically 10–15° apart. This results in the brain receiving two slightly different images which, when combined, give the third dimension to the object observed, just as in normal everyday vision.

Most manufacturers offer very fine microscopes with highly corrected objectives and methods to record an image on film or videotape. Stereomicroscopes are used most frequently as inspection devices; however, to the chemical microscopist, the uses include preliminary sample observation or sample selection, segregation of materials, manipulation of a sample, and preparation such as picking a particle or fiber from a bulk sample and mounting it in a particular orientation on a microscope slide in preparation for more detailed study under the higher magnification of a compound microscope, most usually, a polarized light microscope.

2. Compound Microscopes

The optical system of a simple compound microscope consists of four basic elements: a light source, a condenser, an objective lens, and an eyepiece or ocular.

The light source in present-day microscopes consists of a tungsten filament light bulb or a higher brightness quartz–halogen bulb. Some specialized techniques, however, require more sophisticated light sources. Fluorescence microscopy, for example, requires a lamp which will emit light at a wavelength (365 nm) sufficient to cause the specimen under observation to fluoresce in the visible light range. The best sources for fluorescence are mercury vapor arc lamps with a set of filters to remove unwanted wavelengths and heat. The condenser is used primarily to collect the light from the source and concentrate the light upon the specimen plane in order to uniformly illuminate the sample. Condensers have other uses which are discussed below.

The objective lens is the most important optical element in the compound microscope; it gathers the light transmitted through or reflected by the specimen and forms a primary image then further enlarged by the ocular. The objective is usually engraved on its barrel with various numbers and letters; for example, $10 \times$ POL, 0.25 or 40 Ph, 0.65, 0.17. The first, and usually largest number, is the magnification; in the two examples above, this would be $10\times$ and $40\times$. This is the nominal magnification of the intermediate image formed at the focal point of the eyepiece. The letters signify the type of objective in use which, in the examples, should be POL for polarizing and PH for phase contrast. Others could be HI for homogeneous (oil) immersion, EPI for episcopic, PLAN for flat field corrected, APO for apochromatic, as well as combinations of these and others.

The next set of numbers in the examples, 0.25 and 0.65, respectively, is the numerical aperture (NA), a measure of the light-gathering capability of the objective and, therefore, its resolution. Since the purpose of a microscope is not solely to magnify an object but rather to resolve fine details of the object, a higher NA objective is preferable to a low-NA objective. The NA was stated by Ernst Abbe to be related to Snell's law as follows:

$$\mathrm{NA} = n \sin \tfrac{1}{2}\mathrm{AA},$$

where n is the lowest refractive index of any element in the lens/specimen system and AA is the angular aperture (the angle between the two most divergent beams entering the front element of the objective).

Diffraction theory states that if an object made up of fine details is illuminated by a beam of light, diffraction maxima will be formed on either side of the perpendicular incident ray. The finer the detail, the larger the diffraction angle; consequently, a wider AA (i.e., higher NA) would be necessary to capture the diffracted rays. In addition to the NA, microscope resolution depends on a number of other factors, such as chromatic and spherical aberration, coma, and astigmatism, any one of which can adversely affect the image quality. For best results each objective must be used with the proper thickness and refractive index of all materials between the object and the objective front lens. Very important is the thickness of the coverslip, especially for high-power dry objectives. The 0.17 on the $40\times$ objective, for example, signifies the thickness of the coverslip for which the objective has been corrected for spherical and chromatic aberrations. Objectives are available with few or all of these problems corrected to varying degrees and at appropriate costs.

The eyepieces, or oculars, are the final stage of magnification. Oculars magnify the image formed by the objective and in some cases supplement the corrections of the objective. Oculars range in power from $5\times$ up to $30\times$; however, increasing the magnification of the objective and the ocular beyond a certain point does nothing to increase resolution, rather it delivers "empty magnification" due to the limit of resolution governed by the wavelength of light used (usually white light). The generally accepted rule for maximum useful magnification (MUM) is $1000\times$ the NA of the objective used. Thus, a 95×1.3NA objective would not be able to show any more detail with the $25\times$ ocular (total magnification $= 2375\times$) than with a $15\times$ ocular ($1425\times$).

Numerous variations of the compound microscope are available, the most basic being a standard biological microscope and the most complex being various types of interference microscopes. Microscopists often will have a universal or modular type of microscope which can be altered by adding intermediate attachments or substage components depending on the type of analysis to be performed. The polarized light microscope is best for chemical microscopy. Under some conditions, most frequently in a crime lab where two similar samples must be compared for possible common origin, two identical microscopes can be optically connected to a single viewing head. This is known as a comparison microscope; it is used for fiber, hair, tool mark, bullet, and cartridge-case comparisons.

3. Spectrometers

The microscope has not been immune to the revolution in laboratory instrumentation whereby most current analytical tools are now computer aided, if not computer operated. New interest in the analysis of very small amounts of materials has prompted individuals and manufacturers alike to design and build instrumentation for attachment to

or including a microscope as an integral part of a larger system. The ultraviolet, infrared, and other radiating sources beyond the spectrum visible to the eye can all yield information about a sample; however, translating this information into something like a spectrum requires an "eye" which can "see" these wavelengths (i.e., a spectrometer). Once the spectrometer collects the information, it can be digitized and stored in a computer for retrieval to display on a CRT, provide a hard-copy printout, or be manipulated in a variety of ways.

Spectrophotometric systems are helpful in matching certain types of unknown materials to standards which can be stored in the computer and are most useful in identifying complex organic or inorganic substances and in quality control applications.

An ultraviolet–visible (UV-VIS) spectrophotometer incorporating a microscope to image small particles or areas of a sample is very useful for problems involving color matching of samples, such as dye compound identification, fiber matching, ink matching, paint identification, or any other problem where spectral content of the sample is important. This instrumentation is commercially available in both reflectance and transmission modes.

A Fourier transform infrared (FTIR) spectrophotometer goes a bit further than the UV-VIS system, enhancing the identification of small amounts of organic material; however, optics, sample handling, and a good background in infrared spectroscopy are essential for getting the information present in the sample out of the complex spectrum generated. FTIR will show what functional groups are present in an organic substance and identify most anions and cations in inorganic compounds. Numerous standard spectra for materials are available and can be kept on file in the system's computer.

Problems involving absorption of IR wavelengths in the optical glass of standard microscope lens systems made it necessary to construct reflecting microscopes as an integral part of modern FTIR spectrophotometers in both transmission and reflectance modes. A computerized system allows for rapid background removal and comparison to a reference beam so that scan times of less than a minute are now possible and library searches of known materials can be rapidly performed. If one is diligent about sample preparation and handling, quantitative information can be obtained.

Concurrent with the development of microscopical FTIR spectroscopy has been the combination of light microscopy, raman scattering spectroscopy, and the laser: the laser raman microprobe (MOLE). The raman effect is a measure of the change in the frequency of monochromatic light as it illuminates an object. Differing from rayleigh scattering in which most of the light is scattered at its original frequency, raman shifts occur both above and below the wavelengths of the illuminating beam; however, these changes occur only with a magnitude of a few parts per million, and conventional light sources can take days to produce enough shifts to generate a spectrum. A laser now provides enough light at a single wavelength so that a spectrum can be obtained in a matter of minutes. When the collimated laser beam is directed through the objective of a microscope such that the beam will impinge upon the sample after first passing through the objective, it becomes even more collimated. The same objective is then used to collect the raman scattered light and direct it to the spectrophotometer. Any sample which can be viewed with the microscope and is not diatomic, fluorescent, or sensitive to the laser beam can be analyzed with this microprobe.

Raman spectra and infrared spectra are similar in that they are based on molecular bond shifts and both spectra yield similar information. Raman spectra, however, are a great deal more complex in that not only does this spectrum indicate types of bonding that occur on the molecular level, it also indicates molecular positioning, thereby allowing differentiation between different polymorphs of a compound, which is generally not achieved with IR. Presently, the only instrument commercially available is the MOLE, and little in the way of prepackaged standard spectra is available; therefore, the analyst should run a standard against the unknown or, at best, have a good idea of the nature of the unknown first.

B. Particle Beam Systems

When particles or surfaces smaller than the resolution of a light microscope ($\sim 0.15 \mu$m) are encountered, an entirely different radiant source must be utilized. Since the wavelength of electrons is less by a factor of about 10^{-5} than the wavelength of visible light, it follows that a microscope using a beam of electrons as a source might allow visualization of much finer detail (i.e., better resolution). Since electrons are electronegative, they can be focused by means of electromagnetic or electrostatic lenses. Various types of electron beam microscopes have been developed, including the TEM, SEM, STEM, electron probe microanalyzer (EPMA), and the field emission microscope, two of which are discussed: the transmission electron microscope (TEM) and the scanning electron microscope (SEM).

1. Transmission Electron Microscope

The TEM operates exactly as its name implies: a beam of electrons is transmitted through the sample. The illuminating beam of electrons is emitted by a thermionic source, usually a fine tungsten wire electrically heated at high voltages under fairly high vacuum, typically 10^{-5} Torr. The beam then is passed through a series of lenses analogous to the lens system of a compound microscope. When the beam passes through the sample, crystal diffraction by the

sample causes some of the incident beam to be scattered away from the normal and some to be absorbed by the sample. The attenuated beam is then passed through another series of lenses, known as projector lenses, whose strength determines the final magnification. This final image is observed directly on a phosphor-coated screen. Resolution is dependent on wavelength, which, in turn, is a function of the accelerating voltage. A beam of short wavelength corresponding to 100 or more kV yields better resolution. Commercial instruments are available with accelerating voltages, typically of 100 to 200 kV and some as high as 10,000 kV. It is not uncommon to have resolution on the order of 2–3 Å and magnifications above 10,000,000×. Chemical information which can be furnished directly by the TEM is a diffraction pattern observed at the back focal plane of the objective lens of the microscope. This pattern is identical in nature to that generated in an X-ray diffraction camera and it is possible to identify crystalline materials in this manner. TEM is frequently used to analyze asbestos collected from water or air.

Sample preparation for TEM is an art, and extreme care must be taken during sample manipulation to avoid cross-contamination. Bulk samples must either be crushed or be sectioned with an ultramicrotome to get samples no thicker than a few hundred Ångstroms. Generally with crystalline preparations, the sample should merely be crushed finely and dispersed on a thin carbon film supported on a 3-mm grid.

2. Scanning Electron Microscope

In contrast to the TEM, the SEM can easily observe nearly any reasonably sized sample and furnish the chemical microscopist with much information with little effort. Primary signals generated in an SEM are secondary electrons (SE), backscattered electrons (BSE), characteristic X rays, X-ray continuum, Auger electrons (AE), low-loss electrons (LLE), electron energy loss (EEL), and transmitted electrons. Of value to the chemical microscopist are primarily the SE, BSE, and X-ray spectrum.

The SEM is related to the TEM only by virtue of the fact that it also employs a beam of focused electrons to bombard a sample and generate an image; however, once the beam of electrons passes through the condenser lenses, the similarity ends. After the beam has been passed through a final aperture, a set of scanning coils deflects the beam at various rates across the sample.

The striking three-dimensional images which have popularized the SEM are the result of secondary electrons ejected from the surface of the sample by the bombarding primary electron beam. The SE image is collected by a scintillator or phosphor-coated light pipe which transmits the signal to a photomultiplier and finally to a viewing CRT, which scans simultaneously with the primary electron beam. Since the area scanned is quite small in relation to the area of the viewing CRT, magnification is thus achieved and can be varied by changing the size of the area scanned across the sample. Resolution and magnification in the SEM are not as good as with the TEM: commercial SEMs now routinely work at 60-Å resolution and magnifications of 500,000× are achievable although the theoretical MUM is ∼20,000×. Most work on an SEM, however, is below 5000×, and very high resolution is rarely important. The lower resolution of the SEM is due partly to lower accelerating voltages in the SEM (0.5–40 kV), but mostly to scattering effects of various types of radiation occurring below the surface of the sample with electrons and X-rays emerging from an area larger in diameter than that of the primary beam. Varying the accelerating voltage also changes the depth to which the primary beam penetrates the sample, usually from 1 to 15 μm.

The SEM can reveal much chemical information about a sample; the three-dimensional images can aid in the study of crystalline material and sample morphology and BSE images can often be used to determine areas of varying composition. The signals of most import to the chemical microscopist, however, are X-rays since their wavelengths or energies can be measured, thus identifying the individual chemical elements in the sample.

3. X-ray Spectroscopy

Both the SEM and the TEM generate characteristic X-rays due to beam/sample interactions. When an electron impacts with an atom, probabilities are that a collision with an inner-shell electron will result. If this inner-shell electron is ejected from the atom, an energy deficiency exists within the atom, and an adjacent shell electron will drop into the space vacated by the ejected electron. When this transfer occurs, X-rays of energies and wavelengths specific to the atomic number of the atom are emitted. These wavelengths and energies can be detected and are, therefore, extremely helpful in terms of chemical element identification. In some cases, compound identification can be enhanced by quantification.

Two different systems presently are utilized to detect and measure X-ray generation in electron beam instruments: wavelength dispersive spectrometers (WDS) and energy dispersive spectrometers (EDS).

Energy dispersive systems are more likely to be incorporated with a TEM or SEM, and wavelength systems, or crystal spectrometers, are more commonly the basis for an analytical TEM.

The wavelength dispersive X-ray analyzer is more difficult to operate and requires a higher energy electron beam. Analysis time can be lengthy (minutes rather than seconds for EDS) and multiple detectors must be used if more than a few elements are to be analyzed. The positive aspects

of WDS are a better than 10-fold increase in resolution (<10 eV), very low background, and analysis of light elements as low as boron.

The EDS is simple to operate and easy to maintain. With most samples, preselected instrument parameters and detector position rarely need be changed. Of course, each sample presents a unique set of aspects which must be taken into account prior to analysis. Most qualitative tests are quite simple and results are obtained in a matter of seconds for all elements above sodium in atomic number. The major drawback of EDS is the poor spectral resolution (145–180 eV), which can cause peak overlaps and difficulty in analysis. For example, a sample containing both Al (1486 eV) and Br (1480 eV) would have severe overlap in that region of the spectrum if the unknown were analyzed at low accelerating voltage (kV). If the voltage is raised, and if Br is present, additional bromine lines would appear at 11,907 and 13,287 eV, confirming the presence of Br, since Al will display only the two lower energy lines. In WDS, this would not be seen as a problem because the two low-energy lines would not overlap but be clearly separated. Another drawback of EDS is the relative difficulty in accurately and routinely analyzing elements below atomic number 11, making carbonate or oxide analysis impossible. This is primarily due to detector design; however, special detectors are currently available which circumvent some of these problems.

Almost all X-ray analyzers are powerful computer-based systems which allow for extensive data storage, rapid access of files, and simple quantitative analysis of samples. The combining of these spectrometers with electron beam microscopes has allowed routine detection of $<10^{-18}$ g of some elements, and since these areas are also imaged by the instrumentation, this elemental analysis can be related to microstructure.

4. Auger/ESCA Spectroscopy

Frequently, analyses must be performed on samples or surfaces which (1) cannot be imaged or analyzed due to small sample size, (2) would be subject to damage by a high voltage (>10 kV) electron beam, or (3) can be imaged only under low kV, and therefore little X-ray information is produced. These samples are commonplace in organic analysis, metallurgy, adhesives and bonding technology, and crystal-growth technology. Recent advances in high-vacuum technology have led to microanalytical tools which can now help solve problems in those fields: Auger (pronounced "ō-zhā") electron spectroscopy (AES) and electron spectroscopy for chemical analysis (ESCA).

The energy deficiency which exists in an atom which has lost an inner-shell electron due to bombardment by an electron beam can be equalized by replacement with an adjacent-shell electron. This energy transfer releases X-radiation or can result in the ejection of an Auger electron.

These electrons are of fairly high energy (50–2000 eV) and the energies are specific to elements present in the sample, additionally, since the amplitude of the Auger peaks are a function of amounts of material present, results can be quantified.

The instrumentation utilizes an electron beam of low accelerating voltages (<5 kV) and large diameter which results in very low penetration into the sample (10–100 Å). By adding an ion gun to the system which, by atomic bombardment of the area under analysis, can remove layers of the structure, a layer-by-layer profile of the chemistry of the material under investigation can be obtained. Because of the high sensitivity of this system, an ultrahigh vacuum of $>10^{-9}$ Torr is desirable and often mandatory.

ESCA does not use an electron beam to excite the sample; rather, it illuminates the sample with a band of X-rays (photons). Upon impact with an atom, these photons will give up some of their energy to electrons which become photoelectrons and are ejected from the atom. An electron spectrometer placed near the illuminated area then detects and measures the intensity or kinetic energy of these photoelectrons. Since the energy of the X-rays is known, the kinetic energy of the photoelectrons can be subtracted from the X-ray energy to yield the binding energy of the ejected electron. This binding energy is unique to each different chemical configuration relative to molecular structure (e.g., to differentiate between a nitrate and a nitrite). ESCA is frequently the instrument of choice in matters of organic surface analysis; however, ESCA, like AES, is a surface probe (10–50 Å) and as such requires care during specimen preparation in order to avoid contaminating the surfaces. Additionally, ultrahigh vacuum systems must be used with ESCA in order to prevent the sample surface from becoming contaminated by gases and vapors in the system.

Until very recently, it had not been possible to analyze samples smaller than a few square millimeters with ESCA, since the illuminating X-ray beam could not be collimated into a fine enough spot. Currently, commercial instruments have become available which can illuminate a 100-μm^2 area to allow analysis of similarly sized samples.

III. POLARIZED LIGHT MICROSCOPY

A. Significance of Polarized Light Microscopy to Chemical Microscopy

The polarized light microscope (PLM) is rather unique in that no other instrument can reveal as much about sample structure and therefore its chemical nature. Numerous

physical properties relating to the chemistry of a material can be observed directly with PLM and quantified in little time by an experienced microscopist. Recognition of this fact can be seen in the recent increase in both sales of PLMs and in the increasing number of students attending industrial and chemical microscopy courses given at a few universities and research institutes.

B. Instrumentation and Accessories

The polarized light microscope is a highly specialized tool quite unlike a medical or biological microscope. Their only similarity is in presenting a magnified image of the sample to the user.

The basic element in any microscope is the stand. The PLM stand should be quite sturdy and be capable of accepting a variety of auxiliary components. Most current microscopes have illuminators built into the base, and these should be able to exhibit good Köhler illumination (see the final paragraph in this section).

Immediately after the illuminator and usually incorporated into the condenser system is a polarizer. This allows light of only one vibration direction to pass through the sample. The polarizer should be oriented such that the plane of vibration of the transmitted light lies in the east/west direction.

Further along the optical path, between the objective and the ocular, lies a second polarizer, known as an analyzer. When the analyzer is inserted into the light path and rotated so that its vibration direction is perpendicular to the polarizer, all light transmitted by the polarizer is absorbed. The analyzer is freely rotatable and is graduated by a scale, measurable in degrees of rotation. The analyzer should be easily removable from the light path by a sliding mechanism. Slots are usually incorporated in the analyzer tube for inserting various types of compensators. A Bertrand lens allows visualization of interference figures and checking the illumination. One ocular should be fitted with a crossline reticle and another with a micrometer scale, and both should be focusable, and the interpupillary distance should be easily varied if the microscope is fitted with a binocular head.

The stage of a PLM is usually circular, freely rotatable, and centerable with respect to the instrument axis. One British design has simultaneously rotating polars (polarizers) and a fixed stage. Either arrangement allows a sample to be observed through all orientations with respect to the vibration planes of the polarizer and analyzer. Stage clips or an attachable mechanical stage may be desirable for certain applications; however, for chemical work, they can be a detriment to simple and rapid sample changes or manipulation.

A rotating nosepiece, capable of mounting four or more objectives, is generally supplied. Objectives should be polarizing objectives (i.e., strain-free) and should be centerable in the nosepiece or individually centerable in individual mounts as well as parfocal. Use of objectives not designed for polarized light may impart unwanted polarization colors to a sample due to strain introduced into the glass lens during manufacture.

The condenser should also be strain free, and the NA should be near the highest objective NA to be used. The condenser should be centerable in its mount in order to aid in bringing the optical system into perfect concentricity. Often, the sample may require a different type of illumination, such as phase contrast or dark field. Provision should be made for simple changeover to different condensers, which must be used for these types of transmitted illumination techniques.

Various compensators such as a quartz wedge, a "first-order red" plate, or a quarterwave plate can be introduced into the bodytube. Compensators aid in the determination of degree of retardation or birefringence and in determining the orientation of the high refractive index.

A desirable addition would be an incident-light illuminator for observing opaque substances or reflected light phenomena. Also, a photomicrographic attachment is useful for recording various effects due to illumination, for comparison of similar samples by use of photomicrographs, or to provide a record for inclusion in a notebook or report.

Dynamic studies such as crystal growth, microchemical reactions, or changes due to thermal effects can be best recorded with a video camera and recorder. These can be easily adapted to any microscope having the capability to accept photomicrographic systems.

Of foremost import regarding the use of *any* light microscope is the ability to set up the illumination system correctly. "Correctly" means Köhler illumination, whereby the images presented have good sharpness, contrast, and are evenly illuminated over the entire field for all objectives. Most microscope users (as distinguished from microscopists) simply turn on the lamp and go about business as usual. If a few minutes were spent adjusting for Köhler illumination, however, optimum resolution and contrast would be ensured at all magnifications and for all samples. Detailed steps for these operations are outlined in any of the referenced McCrone publications.

C. Observable Optical Phenomena

An experienced microscopist can often identify many frequently encountered substances at sight. Unfortunately, the microscopist is less likely to be believed if he/she says something is, for example, cornstarch rather than

saying, "These particles are approximately 5–15 μm in size, are colorless, transparent, and rounded polyhedra. The refractive index is about 1.530 and each grain contains a central nonspherical air bubble. The surface is smooth, there is no agglomeration, and when observed between crossed polars, the particles appear gray to white with a distinct black cross." To further substantiate the identification, a dilute aqueous drop of I-KI (iodine–potassium iodide) added to the particles will color starch blue to nearly black. This entire identification can be performed in less than 15 sec, additionally 10 different bits of information were obtained by simple observation. Obviously all samples are not so readily characterized; nevertheless, the techniques remain the same and much additional information about processing and source can be obtained.

Preliminary examination under a stereomicroscope can reveal whether a substance is crystalline or amorphous and can aid in determining which area of a bulk sample is representative of the problem at hand. A sample can also be more easily manipulated and mounted under the stereoscope. If the sample is large enough, it is best to try to break the sample into a few smaller particles so more than one prep is available for testing. Preparing a mount generally means placing a sample on a microscope slide and immersing it in a suitable medium, preferably with a refractive index of 1.660. The sample is then ready for observation by PLM.

Once the sample is focused and observed in plane polarized light, a number of simple observations are possible. Particle size can be measured with a calibrated measuring eyepiece reticle or a filar micrometer eyepiece. Once sizes are established, a size distribution of particles can be estimated as can the percentage of various types of particles if the sample is a mixture.

The general appearance of the particle(s) can also be noted. Transparency or opacity should be noted as should shape and surface texture. Unless one is going to be addressing specialized specific types of particles all of the time (e.g., fibers or mineral samples), it is best not to become too specific about shape or texture nomenclature. Easily conveyed terms such as flakes, fibrous, granular or cracked, smooth or pitted usually suffice to describe the general overall shape and surface texture of the sample. If analysis of the sample by optical crystallography becomes necessary, then terms such as orthorhombic bipyramids, or monoclinic prisms, if correctly applied, are necessary to classify a crystal. Observations of surfaces are generally aided with the use of an incident illuminator to light the top surface of the sample. Color of the sample should also be observed with both incident and transmitted light. Pleochroism can be determined while the sample is illuminated by transmitted plane polarized light. Pleochroic materials exhibit different colors with changes in orientation relative to the plane of light vibration. Simply stated, if the color of the particle changes as the stage is rotated, the particle is pleochroic, thus also anisotropic.

The relative hardness of a material can also be tested by simply pressing down gently on the coverslip within the vicinity of the particle with a fine tungsten needle. Brittle particles will easily fracture or flake apart; a particle which spreads out and then recovers its original shape will indicate some polymerized or rubberized material; very hard particles may do nothing. Magnetic properties can easily be tested by passing a small magnet past the objective while observing its movement or lack thereof.

One of the most helpful keys in morphological analysis of microscopic samples is knowing the refractive index of the sample. The reason for selecting a mounting medium with a refractive index of 1.660 is that most substances have refractive indices on either side of 1.660 (the best mountant for this purpose has proved to be Aroclor, which is available in various viscosities). For example, glass ($n_D = 1.515$) and the mineral corundum Al_2O_3 ($n_D = > 1.7$) appear to be similar under the microscope but if the sample has been mounted in Aroclor, the relative refractive index of the particle determined by the "Becke line" test will quickly differentiate the two. The Becke line is the bright ring or halo visible around the boundary of the particle visible as the plane of focus is raised or lowered. The halo will always move from the medium of lower refractive index to the medium of higher refractive index as the plane of focus is raised. Thus, if focus is raised on the particle in question and the halo moves from the outer perimeter of the particle to the inner perimeter, the refractive index of the particle is higher than the mountant; in this case, the unknown could be corundum. At this point, simply inserting the analyzer at the crossed polars position (90° to polarizer) would confirm this fact; since glass is isotropic and corundum is anisotropic, the glass would be lost to sight and the corundum would appear bright against a black field.

Once an approximation of the refractive index of the particle is determined, an exact match can be determined by plotting a dispersion staining curve. Dispersion is the variation in refractive index with wavelength of light. By using (1) a suitable mounting liquid, (2) a dispersion staining objective, (3) a hot stage, and (4) graph paper, a curve can be obtained by plotting the wavelength of the particle's observed dispersion staining colors (from charts) versus the variation in the refractive index of the medium, which changes with temperature. Two pieces of data are obtained from this test; the experimental refractive index at which the particle is no longer visible is established as the refractive index of the particle and the curve obtained can be compared to known literature data to obtain a match,

therefore, an identity. Dispersion staining is also a very useful technique for determining the presence of a known substance in a mixture; it would be possible to estimate the percentage of sucrose in a cake mix, determine drug purity, or determine asbestos content in an insulation sample very rapidly with appropriate dispersion staining methods.

Observation of samples with the analyzer now inserted reveals other information about a material. With the analyzer set to the crossed polars position (90° to polarizer), particles which remain dark while rotating the stage are termed isotropic and exhibit only a single refractive index. All crystals in the cubic system, some polymers, and unstrained glasses are isotropic and will not affect the light passing through them aside from refraction. The vast majority of samples observed will be anisotropic and will appear white or colored under crossed polars.

Anistropic or doubly refractive substances exhibit more than one refractive index and are termed birefringent materials. When plane polarized light enters a birefringent crystal, the particular molecular lattice structure of that crystal will split the light into two *different* but perpendicular vibrational directions termed component vibrations, each traveling at different velocities, therefore indicating different indices of refraction. By rotating the stage, the particle will change from bright to dark, or extinction, every 90°. The new vibration axes of each of the component vibrations can be measured by aligning a crystal face parallel to the crossline in the ocular, noting the angle indicated on the stage vernier and now rotating the stage until the crystal loses all color. At this point, the stage angle is read again and the difference between the two readings is the extinction angle of the crystal.

Since the two components travel at different rates through the crystal, they become separated as the slow component is retarded behind the fast component. This difference in path length between the two, termed the retardation, can be estimated and measured in nanometers. Retardation is observed as polarization colors resulting from destructive interference which occurs as the two components pass through the analyzer. Retardation is estimated by use of variable compensators such as the quartz wedge or Berek compensator. Sample thickness is related to retardation and birefringence as follows: $r = 1000tB$, where r is retardation in nanometers, t is thickness in microns, and B is birefringence $(n_2 - n_1)$; therefore, if any two of the above quantities are known, the third may be obtained. These data, in conjunction with a Michel–Lévy birefringence chart, are also helpful in identifying substances. The colors on the chart follow the same order as one sees as the variable compensator is inserted (or for the Berek, rotated). The color sequence is known as the Newton series. Once familiar with the colors in the series, it is relatively simple to estimate the retardation of a sample. Some materials show very low order polarization colors making quantitation difficult. In such cases, a first-order red-plate compensator is used to shift the colors higher into the Newton series and give colors (blue or yellow) to gray and white particles.

If necessary at this point, additional optical crystallographic data may be obtained. Suffice it to say, however, that most materials may be characterized by polarized light microscopy: interference figures, crystal systems, symmetry, form, habit, crystallographic axes, and signs of elongation. These data added to information already obtained should give positive identification (if that substance has been characterized previously).

D. Microchemical Tests

On occasion, a sample may be more rapidly identified, or the suspected identity may be confirmed, by microchemical tests. These tests are scaled-down versions of known chemical reactions and can be performed on inorganic or organic materials. The previously mentioned test for starch is one such example.

One should not attempt microchemical tests without first conducting a morphological examination of the sample as previously outlined. This examination should limit the sample to a few possibilities, which may be reduced to a final known with careful treatment and technique.

Chemical tests on a microscale are based on many well-known reactions; color, precipitation, or other types of reactions which occur and can be observed on a macroscale can all be miniaturized to be performed on individual particles of samples. Sensitivities of these tests depend on a number of criteria: the lowest concentration of material in a mixture which will always yield a positive test, the absolute smallest mass or concentration of a single material which always yields a positive test, and varying chemical phenomena not always controllable by the analyst (i.e, solubility, catalysis, rate of reaction, etc.).

Above all, meticulous technique is imperative as is common sense with regard to approaching the sample from a *chemical* point of view. There can be no predetermined standard procedure used when approaching an analysis, although actual tests are basically cookbook type. Most importantly, a preformulated, well-thought-out idea with regard to what the test is intended to illustrate should be the first consideration.

Numerous reagents for testing exist, however, for most determinations, no more than a few dozen should be necessary. An excellent set of reagents is available from Cargille Labs (Cedar Grove, NJ), which should cover most inorganic tests, and numerous others are listed in the various references. Most common solvents should also be available.

A prime advantage of microchemical tests is the low cost of testing. Aside from the microscope, cost for all ancillary equipment, including reagents, should be a few hundred dollars; the reagents are used in amounts of only a few crystals per test; thus, the initial purchase may last many years. Associated equipment necessary may include a microburner or hot plate, thin glass rods, a length of platinum wire, microcapillary tubes, and scrupulously clean slides and coverslips.

Since the sensitivity of most tests is in the subnanogram range, it is frequently possible to perform a range of tests on a single particle. A first step might be to observe the sample's reaction to heat. Does it melt at a very low temperature or not at all? This would be a good indicator of a sample being organic or not and would eliminate numerous possibilities. What solvents affect the sample? Solubility tests can be easily performed with only nanoliters of solvent as follows: (1) place the particle on a coverslip, (2) place a section of glass tubing (10–15 mm in diameter and 10 mm high) upright on a glass slide, (3) carefully invert the coverslip with particle and set it on top of the well, and (4) apply a drop of solvent to the base of the well. Capillarity will cause the solvent to flow into the chamber and vapors will be seen to affect the particle if it is soluble. Slightly heating the slide near the well will increase the vapor pressure when necessary. Microchemical tests can often be made on single particles in this same cell by placing a reagent nearby on the coverslip. Moisture pickup gives two droplets, which then coalesce and react to give the test.

Most tests require that the sample be brought into solution before being tested. Occasionally, a sample may not be readily soluble in most reagents or only very slowly soluble. These may need to be "opened up" by means of high temperature fusion in a flux such as sodium carbonate. This can also be performed easily on a microscale (i.e., a fine platinum loop) with appropriate fluxes and technique.

Once the sample has been brought into solution, reagents for the specific test(s) may be introduced and, in most cases, the resulting crystals, precipitate, or color will be immediately apparent. Reagents may be introduced in a number of ways and the specific method recommended should be employed. Most frequently, "negative" tests result from inappropriate application of the reagent or hurried, inexact technique. If a negative test results, the sample will still be in solution and through careful treatment, can still be used for additional testing.

Another advantage of microchemical tests is that not only can cations be detected, but anions can be quite specifically identified as well. Silver nitrate or barium chloride as reagents can pinpoint more than 50 different anions. This is not as easily determined with more exotic electron beam instruments.

As an example of how quickly these tests can be performed, a few examples follow:

1. A sample is found to be opaque, noncrystalline, and highly reflective, with a silvery-gray metallic appearance. Solubility testing indicates the sample to be slowly attacked by dilute HCl with effervescence. An iron (ferric) containing metal is suspected; therefore, the sample is treated with HNO_3 to ensure obtaining ferric ions and a crystal of potassium ferrocyanide is added to the test drop. If the resulting precipitate is dark blue (prussian blue) in color, the test is positive for ferric iron.
2. Fibers found in the trunk of a car submitted for analysis are observed to have small reddish-brown deposits clinging to the outside. The fiber, found to be nylon by use of a quartz wedge and Michel–Lévy chart, readily releases the material in water. The amorphous appearance and reddish color indicates iron or possibly blood. A microchemical test for iron is negative; therefore, a test for blood is attempted. The reagent (Takayama reagent) is a mixture of water, 10% glucose, 10% sodium hydroxide, and pyridine. When a drop is added to the particles, heated gently but not to boiling, and allowed to stand for several minutes, red, lightning-shaped crystals result, specific for blood.
3. A colored compound in a red rubber is thought to contain iron, copper, or cadmium. The organics in the rubber are first removed in a flame, and the residual ash is placed on a slide. After treating the material with HNO_3 and diluting it, a crystal of potassium mercuric thiocyanate is added to the drop and observed. Iron will give no crystals, cadmium will yield very specific colorless crystals pointed at one end, and copper will yield greenish-yellow boat-shaped crystals or dentritic "trees." A specific test for iron could be run (e.g., KSCN or the above-mentioned Prussian blue test).

The only way to become adept at chemical microscopy is through hands-on daily practice. None of these techniques are push-button routine procedures but, once mastered, the morphological identification of materials, combined with microchemical testing, can be a valuable asset to any mateirals personnel with the side benefit of observing some of the most dramatic and delightful images in nature.

SEE ALSO THE FOLLOWING ARTICLES

AUGER ELECTRON SPECTROSCOPY • ELECTRON MICROPROBE ANALYSIS OF MINERALS • IMAGING OPTICS • PARTICLE SIZE ANALYSIS • POSITRON MICROSCOPY • SCANNING ELECTRON MICROSCOPY

• SCANNING PROBE MICROSCOPY • TRANSMISSION ELECTRON MICROSCOPY • X-RAY ANALYSIS • X-RAY PHOTOELECTRON SPECTROSCOPY

BIBLIOGRAPHY

Benedetti, P. A. A. (1964). "Identification of materials via Physical Properties, Chemical Tests, and Microscopy," Springer-Verlag, New York.

Chamot, E. M., and Mason, C. W. (1940). "Handbook of Chemical Microscopy," Vol. II, 2nd ed., Wiley, New York.

Feigl, F. (1949). "Chemistry of Specific, Selective, and Sensitive Reactions," Academic Press, New York.

Feigl, F. (1972). "Spot Tests in Inorganic Analysis," 6th ed., Elsevier, New York.

Feigl, F. (1975). "Spot Tests in Organic Analysis," 7th ed., Elsevier, New York.

Fulton, C. C. (1969). "Modern Microcrystal Tests for Drugs, the Identification of Organic Compounds by Microcrystalloscopic Chemistry," Wiley (Interscience), New York.

McCrone, W. C. (1980). "The Asbestos Particle Atlas," Ann Arbor Science, Ann Arbor, MI. (Available from MAC, 2506 S. Michigan Ave., Chicago, IL 60616.)

McCrone, W. C., and Delly, J. G. (1972, 1979). "The Particle Atlas," Vols. I–VI, Ann Arbor Science, Ann Arbor, MI. (Available from MAC, 2506 S. Michigan Ave., Chicago, IL 60616.)

McCrone, W. C., McCrone. L. B., and Delly, J. G. (1979). "Polarized Light Microscopy," Ann Arbor Science, Ann Arbor, MI. (Available from MAC, 2506 S. Michigan Ave., Chicago, IL 60616.)

Mason, C. W. (1983). "Handbook of Chemical Microscopy," 4th ed., Vol. I, Wiley, New York.

Schaffer, H. F. (1953). "Microscopy for Chemists," Dover, New York.

Schneider, F. L. (1964). "Qualitative Organic Microanalysis," Academic Press, New York.

Optical Interferometry

P. Hariharan
University of Sydney

GLOSSARY

Coherence Statistical measure of the similarity of the fields produced by a light wave at two points separated in space and/or time.

Doppler effect Shift in the frequency of a wave observed with a moving source or detector, or when it is reflected or scattered by a moving object.

Incoherent source Source consisting of a large number of individual emitters that radiate independently of each other.

Laser Light source that radiates by stimulated emission. The output from a laser is highly directional and monochromatic.

Localized fringes With an incoherent light source the contrast of the fringes produced by an interferometer is usually a maximum in a particular plane. The fringes are then said to be localized in this plane.

Moiré fringes Relatively coarse fringes produced by the superposition of two fine fringe patterns with slightly different spacings.

Nonlinear optical materials Materials whose refractive index changes with the intensity of the incident light.

Optical fiber Glass fiber surrounded by a transparent sheath with a lower refractive index. If the diameter of the fiber is comparable to the wavelength, light propagates along it in a single guided mode.

Optical path Product of the geometrical path traversed by a light wave and the refractive index of the medium.

Polarizing beamsplitter Optical element that separates light waves polarized in two orthogonal planes.

Quarter-wave plate Device that introduces an optical path difference of a quarter wave-length between two orthogonally polarized waves.

Speckle Granular appearance of a rough surface illuminated by a laser caused by the superposition of scattered light waves with random phase differences.

Visibility Visibility of interference fringes is defined by the relation $V = (I_{max} - I_{min}/(I_{max} + I_{min})$, where I_{max} and I_{min} are the irradiances at adjacent maxima and minima, respectively.

OPTICAL INTERFEROMETRY comprises a range of techniques that use light waves to make extremely accurate measurements. For many years optical interferometry remained a laboratory technique. However, as a result of several recent innovations such as lasers, optical fibers, holography, and the use of digital computers for image processing, optical interferometry has emerged as a very practical tool with many applications.

I. INTERFERENCE OF LIGHT

If two or more waves are superposed, the resultant displacement at any point is the sum of the displacements due to the individual waves. This is the well-known phenomenon of interference. With two waves of equal amplitude it is possible for their effects to cancel each other at some points so that the resultant amplitude at these points is zero.

The colors of an oil slick or a thin film of air enclosed between two glass plates are due to the interference of light waves. To observe such interference patterns (fringes), the interfering waves must have exactly the same frequency. This normally implies that they must be derived from the same light source. In addition, to maximize the visibility of the fringes, the polarization of the interfering light waves must be the same.

Only a few interference fringes can be seen with white light because as the optical path difference increases, the phase difference between the interfering light waves differs for different wavelengths. However, interference fringes can be seen with much larger optical path differences if light with a very narrow spectral bandwidth is used.

Optical interferometers split the light from a suitable light source into two or more parts that transverse separate paths before they are recombined. Common types of interferometers are the Michelson, the Mach–Zehnder, and the Sagnac, which use two-beam interference, and the Fabry–Perot, which uses multiple-beam interference. Applications of optical interferometry include accurate measurements of distances, displacements, and vibrations, tests of optical systems, studies of gas flows and plasmas, microscopy, measurements of temperature, pressure, electrical, and magnetic fields, rotation sensing, and even the determination of the angular diameters of stars.

II. MEASUREMENT OF LENGTH

One of the earliest applications of optical interferometry was in length measurements where it led to the replacement of the meter bar by an optical wavelength as the practical standard of length. Several lasers are now available that emit highly monochromatic light whose wavelength has been measured extremely accurately. With such a laser, optical interferometry can be used for very accurate measurements of distances of a hundred meters or more.

A. Absolute Measurements of Length

Electronic fringe counting is now widely used for length interferometry. In the Hewlett–Packard interferometer a helium–neon laser is forced to oscillate simultaneously at two frequencies separated by a constant difference of about 2 MHz, by applying an axial magnetic field. As shown in Fig. 1, these two waves, which are circularly polarized in opposite senses, are converted to orthogonal linear polarizations by a quarter-wave plate. A polarizing beamsplitter reflects one wave to a fixed corner reflector C_1, while the other is transmitted to a movable corner reflector C_2. The returning waves pass through a polarizer, so that the transmitted components interfere and are incident on the detector D_S.

The outputs from D_S and a reference detector D_R go to a differential counter. If the two reflectors are stationary,

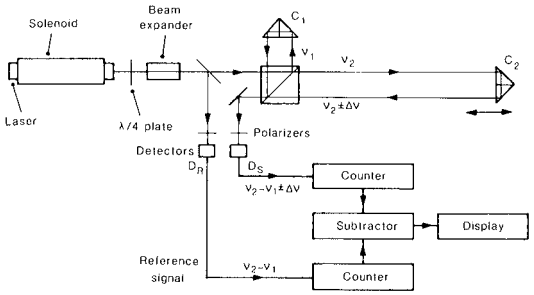


FIGURE 1 Fringe-counting interferometer using a two-frequency laser. [After J. N. Dukes and G. B. Gordon (1970). *Hewlett-Packard J.* **21**(12), 2–8. © Copyright 1986 Hewlett–Packard Company. Reproduced with permission.]

the frequencies of the two outputs are the same, and no net count accumulates. If one of the reflectors is moved, the change in optical path in wavelengths is given by the net count.

Another technique, which can be used if the distance to be measured is known approximately, involves synthetic long-wavelength signals. This technique is based on the fact that if two wavelengths λ_1 and λ_2 are simultaneously incident on a two-beam interferometer, the envelope of the fringes corresponds to the interference pattern that would be obtained with a synthetic wavelength $\lambda_S = \lambda_1\lambda_2/|\lambda_1\lambda_2|$.

The carbon dioxide laser can operate at several wavelengths, which have been measured accurately, and is, therefore, well suited to such measurements. The laser is switched rapidly between two of these wavelengths and the output signal obtained from a detector as one of the interferometer mirrors is moved is squared, low-pass filtered, and processed in a computer to obtain the phase difference. Distances up to 100 m can be measured with an accuracy of one part in 10^7.

Yet another method is to use a semiconductor laser whose frequency is swept linearly with time by controlling the injection current. For an optical path difference D, the two beams reach the detector with a time delay D/c, where c is the speed of light, and they interfere to yield a beat signal with a frequency

$$f = (D/c)(df/dt) \tag{1}$$

where df/dt is the rate at which the laser frequency, is varying with time.

B. Measurements of Very Small Changes in Length

A number of interferometric techniques are also available for accurate measurements of very small changes in length. One method is based on phase compensation. Changes in the output intensity from the interferometer are detected and fed back to a phase modulator in the measurement path so as to hold the output constant. The drive signal to the modulator is then a measure of the changes in the optical path.

Another method involves sinusoidally modulating the phase of the reference beam. Under these conditions the average phase difference between the interfering beams can be determined from a comparison of the amplitudes of the components in the output of the detector at the modulation frequency and at its second harmonic.

A third group of methods is based on heterodyning (light beats). For this a frequency difference is introduced between the two beams in the interferometer, usually by means of a pair of acoustooptic modulators operated at slightly different frequencies. The output from a detector then contains an oscillatory component at the difference frequency whose phase corresponds to the phase difference between the two interfering wave fronts.

Light beats can also be produced by superposing the beams from two lasers operating on the same transition and can be used to measure changes in length very accurately. For this purpose two mirrors are attached to the ends of the specimen to form a Fabry–Perot interferometer. The frequency of a laser is then locked to a transmission peak of the interferometer, so that the wavelength of this slave laser is an integral submultiple of the optical path difference in the interferometer. A displacement of one of the mirrors results in a change in the wavelength of the slave laser and hence in its frequency. These changes are measured to better than one part in 10^8 by mixing the beam from the slave laser at a fast photodiode with the beam from a frequency-stabilized reference laser and measuring the beat frequency.

III. MEASUREMENTS OF VELOCITY AND VIBRATION AMPLITUDE

Light scattered from a moving particle has its frequency shifted by the Doppler effect by an amount proportional to the component of the velocity of the particle along the bisector of the angle between the directions of illumination and observation. With a laser source this frequency shift can be detected by the beats produced either by the scattered light and a reference beam or by the scattered light from two illuminating beams incident at different angles. An initial frequency offset can be used to distinguish between positive and negative flow directions. Laser-Doppler interferometry is now used widely to measure flow velocities. Another industrial application has been for noncontact measurements of the velocity of moving material.

Laser-Doppler techniques can also be used to analyse surface vibrations using an interferometer in which a frequency offset is introduced between the beams by an acoustooptic modulator. The output from a detector then consists of a component at the offset frequency (the carrier) and two sidebands. Vibration amplitudes down to a few thousandths of a nanometer can be determined by a comparison of the amplitudes of the carrier and the sidebands, while the phase of the vibration can be obtained by comparison of the carrier with a reference signal.

IV. OPTICAL TESTING

Another major application of interferometry is in testing optical components and optical systems. The instruments

commonly used for this purpose are the Fizeau and the Twyman–Green interferometers. The Fizeau interferometer is widely used to compare flat surfaces. However, with a laser source it can carry out a much wider range of tests, including tests on concave and convex surfaces.

The output of such an interferometer is a fringe pattern that can be interpreted by an observer quite readily; unfortunately, the process of extracting quantitative data from it is tedious and time consuming. This has led to the use of digital computers for analyzing such fringe patterns.

A. Digital Techniques

A typical digital system for fringe analysis uses a television camera in conjunction with a video frame memory and a minicomputer. Since the fringes only give the magnitude of the errors and not their sign, a tilt is introduced between the interfering wave fronts so that a linear phase gradient is added to the actual phase differences that are being measured.

Much higher accuracy can be obtained by directly measuring the optical path difference between the two interfering wave fronts at an array of points covering the interference pattern. A number of electronic techniques are now available for this purpose.

In one method (phase shifting) the optical path difference between the interfering beams is varied linearly with time, and the output current from a detector located at a point on the fringe pattern is integrated over a number of equal segments covering one period of the sinusoidal output signal. Alternatively, the optical path difference between the interfering wave fronts is changed in equal steps and the corresponding values of the intensity are measured. Three measurements at each point provide enough data to calculate the original phase difference between the wave fronts. Since a photodiode array or a charge-coupled detector array can be used to implement this technique, measurements can be made simultaneously at a very large number of points covering the interference pattern.

The simplest way to generate the phase shifts or phase steps is by mounting one of the mirrors of the interferometer on a piezoelectric transducer to which appropriate voltages are applied. Another way is to use a semiconductor laser whose output wavelength can be changed by varying the injection current. If the optical path difference between the two arms of the interferometer is D, a wavelength change $\Delta\lambda$ results in the introduction of an additional phase difference between the beams,

$$\Delta\varphi \approx 2\pi D\Delta\lambda/\lambda^2 \qquad (2)$$

Figure 2 shows a three-dimensional plot of the errors of a surface produced by an interferometer with a digital phase measurement system. Because of their speed and accuracy such interferometers are now used extensively in the production of high-precision optical components.

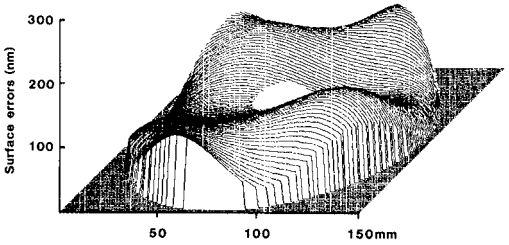


FIGURE 2 Three-dimensional plot of the residual errors of a concave mirror obtained with a digital phase-measuring interferometer.

B. Tests of Aspheric Surfaces

Many optical systems now use aspheric surfaces. The simplest way of testing such a surface is to generate a table of wave front data giving the theoretical deviations of the wave front from the best-fit sphere and to subtract these values from the corresponding measurements. Surfaces with large deviations from a sphere can be tested either by using long-wavelength (infrared) light or by recording phase data with two wave-lengths. These are used to calculate the phase differences between adjacent data points corresponding to a longer synthetic wavelength. The profile of the surface can then be obtained by integrating these differences.

Surfaces with large deviations from a sphere can also be tested with a shearing interferometer in which the interference pattern is produced by superposing different portions of the test wave front. In a lateral shearing interferometer two images of the test wave front are superposed with a small mutual lateral displacement. For a small shear the interference pattern corresponds to the derivative of the wave-front errors and the deviations to be measured are considerably smaller than the errors themselves. Evaluation of the wave-front aberrations is easier with a radial shearing interferometer in which interference takes place between two images of the test wave front of slightly different sizes.

Surfaces with very large deviations from a sphere are best tested with a suitably designed null lens, which converts the wave front leaving the surface under test into an approximately spherical wave front, or with a computer-generated hologram (CGH). Figure 3 is a schematic of a setup using a CGH in conjunction with a Twyman–Green interferometer to test an aspherical mirror. The CGH resembles the interference pattern formed by the wave front from an aspheric surface with the specified profile and a tilted plane wave front and is positioned so that the mirror under test is imaged on to it. The deviation of the surface

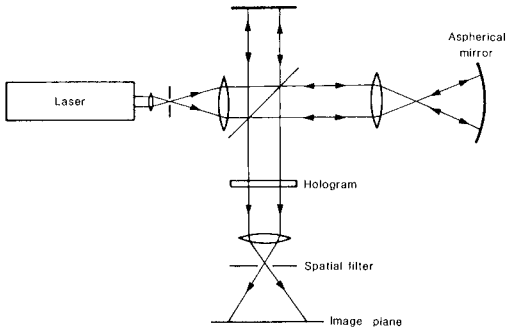


FIGURE 3 Interferometer using a computer-generated hologram to test an aspheric mirror. [From J. C. Wyant and V. P. Bennett (1972). *Appl. Opt.* **11**(12), 2833–2839.]

under test from its specified shape is then given by the moiré pattern formed by the actual interference fringes and the CGH, which is isolated by means of a small aperture placed in the focal plane of the imaging lens.

V. STUDIES OF REFRACTIVE INDEX FIELDS

A significant field of application of optical interferometry has been in studies of diffusion, fluid flow, combustion, and plasmas, where changes in the refractive index can be related to changes in pressure, temperature, or relative concentration of the different components.

The Mach–Zehnder interferometer is commonly used for such studies. It has several advantages for such work: the separation of the two beams can be made as large as desired, the test section is traversed only once, and fringes localized in a plane in the test section can be obtained with an extended incoherent source such as a flash lamp.

Measurements of changes in the optical path difference can now be made extremely rapidly to better than 0.01 wavelength by heterodyne techniques, using either an image-dissector camera to scan the interference pattern, or an array of detectors coupled to individual phase-to-voltage converters.

VI. HOLOGRAPHIC AND SPECKLE INTERFEROMETRY

Holography makes it possible to use interferometry for measurements on objects with rough surfaces. Holographic interferometry is now a powerful tool for nondestructive testing and strain analysis.

Initially, a hologram of the object is recorded by illuminating it with a laser and allowing the light reflected by the object to fall on a high-resolution photographic plate along with a reference beam from the same laser. When the processed hologram is replaced in exactly the same position and illuminated with the same reference beam it reconstructs an image that is superimposed exactly on the object. If a stress is applied to the object, interference between the wave front reconstructed by the hologram and the wave front from the deformed object gives rise to fringes that contour the changes in shape of the object. Weak spots and defects are revealed by local changes in the fringe pattern.

Very accurate measurements of the optical path differences in the interference pattern can be made by the digital phase-stepping technique. The data from three or more such measurements made with different directions of illumination can then be processed to obtain the surface displacements and the principal strains.

Holographic interferometry can also be used for measurements on vibrating objects. One method (time-average holographic interferometry) involves recording a hologram with an exposure long compared to the period of vibration. The reconstructed image is then covered with fringes that can be used to map the vibration amplitude. More accurate measurements can be made using stroboscopic illumination in conjunction with the phase-stepping technique.

A faster, though less accurate, technique for such measurements is speckle interferometry, which involves recording the interference pattern formed between the speckled image of the object when it is illuminated with a laser and a reference beam from the same source. Any change in the shape of the object results in a change in the optical path difference between the two wave fronts and a consequent change in the intensity distribution in the speckled image. Two such speckled images can be recorded electronically and their difference extracted to give fringes similar to those obtained by holographic interferometry.

The digital phase-stepping technique can also be used with electronic speckle pattern interferometry. In this case each speckle is treated as an individual interference pattern, the light from the object having a particular amplitude and phase. If the optical path difference at each such point with respect to the reference beam is measured by the phase-stepping technique before and after the surface moves, the change gives a direct measure of the surface displacement at that point.

VII. INTERFERENCE MICROSCOPY

An important application of optical interferometry is in microscopy. Interference microscopy provides a noncontact method for studies of surface structure when stylus

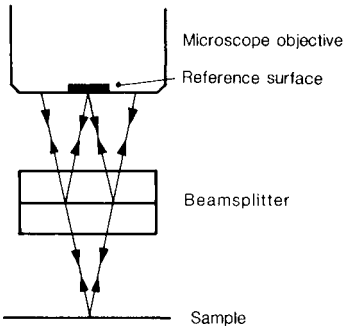


FIGURE 4 The Mirau interferometer.

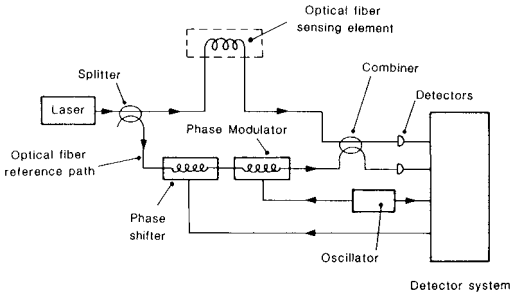


FIGURE 5 Fiber-optic interferometric sensor. [From T. G. Giallorenzi, J. A. Bucaro, A. Dandridge, G. H. Sigel Jr., J. H. Cole, S. C. Rashleigh, and R. G. Priest (1982). *IEEE J. Quant. Electron.* **QE-18**(4), 626–665. © 1982 IEEE. Reproduced with permission.]

profiling cannot be used because of the risk of damage. In the Mirau interferometer shown in Fig. 4, light from the microscope illuminator is incident, through the objective, on a semitransparent mirror. The transmitted beam goes to the test surface, while the reflected beam goes to a reference surface. These two beams are recombined at the same semitransparent mirror and return through the objective. The interference pattern formed in the image plane contours the deviations from flatness of the test surface. Very accurate measurements of surface profiles and estimates of roughness can be made using the digital phase-shifting technique.

Another application of interference microscopy is for studies of transparent living cells that cannot be stained without damaging them. The Nomarski interferometer, which is commonly used for such work, is a shearing interferometer that uses two Wollaston (polarizing) prisms to split and recombine the beams. Two methods of observation are possible. With small isolated objects it is convenient to use a lateral shear larger than the dimensions of the object. Two images of the object are then seen, covered with fringes that contour the phase changes due to the object. With an extended object the shear is made much smaller than the dimensions of the object. The interference pattern then shows the phase gradient.

VIII. INTERFEROMETRIC SENSORS

It is possible to set up interferometers in which the two paths are single-mode optical fibers. Since the optical path length in a fiber changes when it is stretched and is also affected by its temperature, fiber interferometers can be used as sensors for several physical quantities. It is possible to have very long noise-free paths in a small space, so that high sensitivity can be obtained. Figure 5 shows a typical optical setup using optical fiber couplers to divide and recombine the beams and a fiber stretcher to modulate the phase of the reference beam. The output is picked up by a photodetector and measurements are made with either a heterodyne system or a phase-compensating system. Detection schemes involving either a modulated laser source or laser frequency switching have also been used.

Fiber interferometers have been used as sensors for mechanical strains and changes in pressure and temperature. They can also be used for measurements of magnetic and electric fields by bonding the fiber to a suitable magnetostrictive or piezoelectric element.

Another application of fiber interferometers has been in rotation sensing where they have the advantages over gyroscopes of instantaneous response, very small size, and relatively low cost. In this case the two waves traverse a closed multiturn loop made of a single optical fiber in opposite directions. If the loop is rotating with an angular velocity ω about an axis making an angle θ with the normal to the plane of the loop, the phase difference introduced between the two waves is

$$\Delta\varphi = (4\pi\omega L R\cos\theta)/\lambda c \tag{3}$$

where L is the length of the fiber, R is the radius of the loop, λ is the wavelength, and c is the speed of light.

IX. STELLAR INTERFEROMETERS

Even the largest stars have angular diameters of about 0.01 arcsec which is well below the resolution limit of the largest telescopes. However, since a star can be considered as an incoherent circular source, its angular diameter can be calculated from the coherence of the light received from it, which in turn can be obtained from measurements of the visibility of the interference fringes formed by light collected from two points at the ends of a long horizontal base line. For a uniform circular source of angular diameter α, the visibility of the fringes is

$$V = 2J_1\frac{(\pi\alpha B/\lambda)}{(\pi\alpha B/\lambda)} \tag{4}$$

where B is the length of the base line and λ is the wavelength; the visibility of the fringes falls to zero when $B = 1.22\,\lambda/\alpha$.

The first stellar interferometer, which was built by Michelson in 1921, used two mirrors whose spacing could be varied, mounted on a 6-m-long support on the 2.5-m telescope at Mt. Wilson, California. The beams reflected by these mirrors were reflected by two other mirrors to the main telescope mirror that brought them to a common focus, at which the interference fringes were formed.

Several practical problems limited the length of the base line that could be used with this interferometer, the most important being lack of stability of the fringes due to rapid random changes in the two optical paths caused by atmospheric turbulence. Modem electronic techniques have now overcome these problems.

Figure 6 shows the optical system of a stellar interferometer designed to make measurements over base lines up to 1 km. Two coelostats (C) at the ends of the north-south base line send light via a system (OPLC) that equalizes the two optical paths to the beamsplitter (B) where they are combined. Two piezoelectric-actuated tilting mirrors (T) controlled by two quadrant detectors (Q) ensure that the two images of the star are exactly superimposed. Two detectors (D_1 and D_2) measure the total flux in a narrow spectral band in the two interference patterns over a sampling interval of a few seconds. During the next sampling interval an additional phase difference of 90° is introduced between the two beams by two mirrors (S) mounted on piezoelectric translators. The visibility of the fringes can then be obtained from the average value of the square of the difference between the signals from the two detectors.

X. GRAVITATIONAL-WAVE INTERFEROMETERS

Einstein's general theory of relativity predicts the existence of gravitational waves, corresponding to ripples in the curvature of space–time, that propagate through the universe at the speed of light. Since all known forms of matter are nearly transparent to a gravitational wave, and its waveform carries detailed information about its source, gravitational waves could provide a new window into the universe. In particular, observations of gravitational waves should make it possible to test nonlinear aspects of general relativity in regions of strong gravitational fields, such as black holes and colliding neutron stars.

The local distortion of space–time due to a gravitational wave stretches space in one direction normal to the direction of propagation of the wave and shrinks it along the orthogonal direction. This strain could, therefore, be measured by detecting the differential changes in the lengths of the arms of a Michelson interferometer.

Gravitational waves from sources such as coalescing binary neutron stars and black holes are expected to produce strains of the order of 10^{-21}. Several highly sensitive, long-baseline laser interferometers are currently under construction with the objective of detecting gravitational waves.

A typical interferometric gravitational-wave detector, shown schematically in Fig. 7, is basically a Michelson interferometer with Fabry–Perot cavities (up to 4 km long) in each arm. These cavities increase the change in the optical path difference produced by any change in the length of the arms by a factor equal to the number of times the beams traverse the cavities. To reduce sensitivity to fluctuations

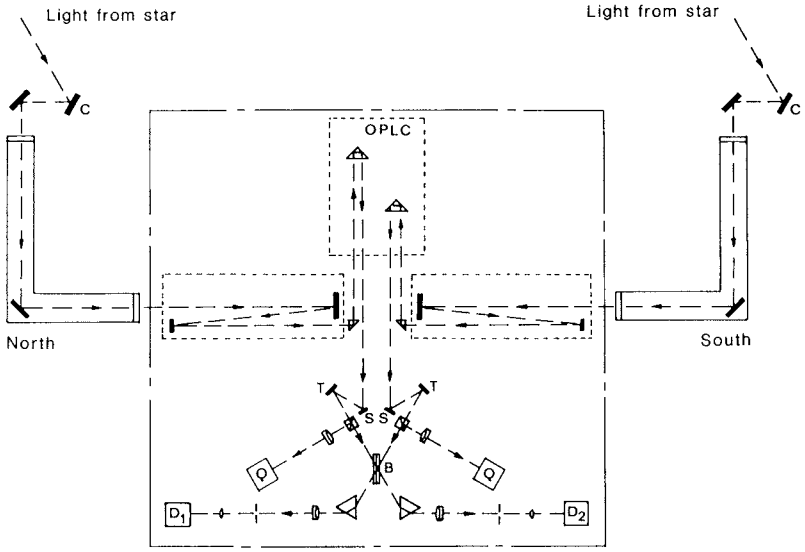


FIGURE 6 A modern stellar interferometer. (Courtesy J. Davis, University of Sydney.)

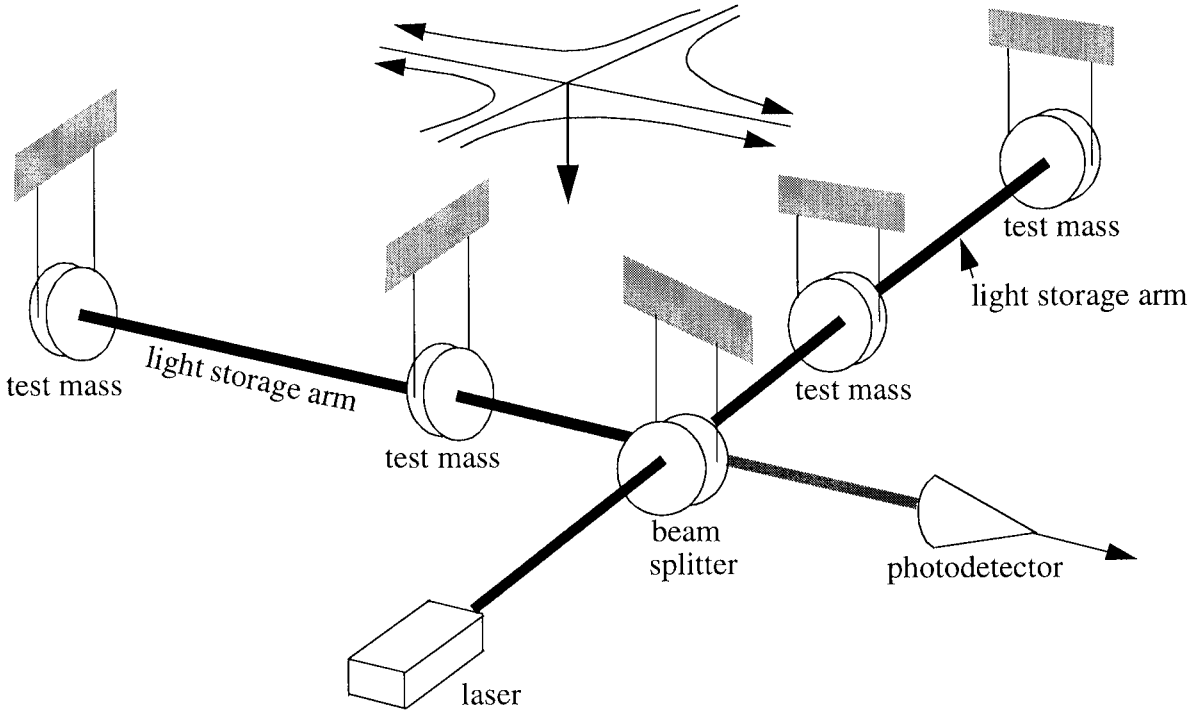


FIGURE 7 Schematic of the optical system of a long-baseline interferometric gravitational-wave detector [From Weiss, R. (1999). *Rev. Mod. Phys.* **71** S187–S196.]

in the input laser power, the antisymmetric or signal extraction port of the Michelson interferometer is held on a dark fringe, resulting in almost all the light returning toward the laser. A further increase in sensitivity can then be obtained by placing a partially transmitting mirror between the laser and the beam splitter. When located at the correct position, this power-recycling mirror reflects most of the light back toward the beam splitter so as to obtain an increase in the effective incident power.

The mirrors of the Fabry–Perot cavities, which constitute four test masses, are suspended from multistage vibration isolators. If L is the separation of the mirrors and h is the strain due to the gravitational wave, the differential change in their separations is then $\Delta L = Lh$, and the resultant change in the phase difference between the beams emerging from the two cavities is

$$\Delta\phi = (2\pi/\lambda)N\,\Delta L = (2\pi/\lambda)NLh, \qquad (5)$$

where λ is the wavelength of the light, and N is the mean number of times the light bounces back and forth in the cavities before exiting.

To achieve the required sensitivity, the laser source has to be frequency and amplitude stabilized, and the entire optical path of the interferometer has to be housed in a vacuum of 10^{-9} torr. Real gravitational waves can be distinguished from instrumental and environmental noise by correlating the outputs of two interferometers at widely separated sites. More information can be extracted from a gravitational wave, including the direction to its source, by combining the outputs from at least three or, preferably, four widely separated sites.

XI. PHASE-CONJUGATE INTERFEROMETERS AND SQUEEZED LIGHT

The availability of nonlinear optical materials has opened up new possibilities in interferometry. One such involves the use of a nonlinear optical material as a phase-conjugate mirror (PCM).

If we consider a distorted wavefront incident on an ordinary mirror, reflection merely inverts the distortion, so that it remains unchanged with respect to the direction of propagation. A diverging wavefront incident on an ordinary mirror is reflected as a diverging wavefront. A PCM, on the other hand, reverses the shape of the wavefront relative to its propagation direction. As a result, a wavefront diverging from a point source becomes, on reflection at a PCM, a converging wavefront moving back to the source. If both the mirrors in a conventional interferometer are replaced with PCMs, a uniform interference field is obtained which is unaffected by misalignment of the mirrors. However, because the PCM responds with a finite delay, the interference pattern shows any transient changes in the optical path. If one mirror is replaced with a PCM, an incoming wavefront can be made to interfere with its

conjugate, and its deviations from a plane wavefront can be obtained directly.

Another interesting possibility is the use of squeezed states of light. Squeezed light can be generated by a number of nonlinear interactions which result in the simultaneous production of pairs of photons. Interferometric measurements with squeezed light can, under appropriate conditions, exhibit smaller residual errors due to quantum noise than measurements made with normal laser light. The use of squeezed light could lead to major improvements in the performance of interferometers for applications requiring the utmost precision, such as the detection of gravitational waves.

SEE ALSO THE FOLLOWING ARTICLES

GRAVITATIONAL WAVE PHYSICS • LASERS, OPTICAL FIBER • LASERS, SEMICONDUCTOR • MICROSCOPY • NONLINEAR OPTICAL PROCESSES • WAVE PHENOMENA

BIBLIOGRAPHY

Born, M., and Wolf, E. (1980). "Principles of Optics," Pergamon, Oxford.

Culshaw, B. (1984). "Optical Fiber Sensing and Signal Processing," Peregrinus, London.

Durst, F., Melling, A., and Whitelaw, J. H. (1976). "Principles and Practice of Laser-Doppler Anemometry," Academic Press, London.

Françon, M., and Mallick, S. (1971). "Polarization Interferometers: Applications in Microscopy and Macroscopy," Wiley (Interscience), New York.

Hariharan, P. (1985). "Optical Interferometry," Academic Press, San Diego.

Hariharan, P. (1987). Interferometry with lasers, *In* "Progress in Optics" (E. Wolf, ed.), Vol. 24. North-Holland, Amsterdam.

Hecht, E., and Zajac, A. (1987). "Optics," Addison-Wesley, Reading, Massachusetts.

Malacara, D. (1978). "Optical Shop Testing," Wiley, New York.

Steel, W. H. (1983). "Interferometry," Cambridge Univ. Press, London and New York.

Vest, C. M. (1979). "Holographic Interferometry," Wiley, New York.

Weiss, R. (1999). "Gravitational radiation." *Rev. Mod. Phys.* **71,** S187–S196.

Radiocarbon Dating

R. E. Taylor
University of California, Riverside

GLOSSARY

Accelerator mass spectrometry (AMS) Method of direct or ion counting of isotopes using particle accelerators as mass spectrometers.

Atomic bomb effect Recent major increase in ^{14}C activity caused by the production of artificial ^{14}C (bomb ^{14}C) due to the detonation of thermonuclear devices in the atmosphere. Also known as the Libby or Nuclear Effect.

Calibrated ^{14}C age Radiocarbon age estimate corrected for secular variation processes.

Conventional ^{14}C age Radiocarbon age estimate expressed in terms of a specified set of assumptions and parameters.

Decay counting Method of inferring ^{14}C concentration by measuring beta decay rate of ^{14}C.

De Vries effects Relatively short-term variations in ^{14}C activities. Also known as "wiggles" or "warps" in the ^{14}C time scale.

Direct counting Method of inferring ^{14}C concentration by directly easuring ^{14}C ions by accelerator mass spectrometry.

Fossil fuel effect Reduction in ^{14}C activity in the 19th and 20th century due to the combustion of fossil fuels. Also known as the Suess and Industrial Effect.

NBS oxalic acid standards Oxalic acid preparations distributed by the U.S. National Standards Bureau of Standards (now National Institutes of Standards and Technology) to serve as contemporary standard for ^{14}C studies. Defines a "zero age" for ^{14}C dating purposes.

Reservoir corrected ^{14}age Radiocarbon age estimate corrected for nonzero age of contemporary materials in a given carbon reservoir.

Secular variation Offsets or deviations from equilibrium natural ^{14}C concentrations due to production rate variations and/or changes in the parameters of the carbon cycle over time.

RADIOCARBON (^{14}C) DATING is an isotopic or nuclear decay method of inferring age for organic materials. The ^{14}C technique is currently the principle time placement or chronometric physical dating method used in late Quaternary geochronology. While the influence of ^{14}C dating has been particularly important in prehistoric archaeological studies, ^{14}C data have made important contributions in geology, geophysics, hydrology, and oceanography through both dating and tracer studies. Radiocarbon measurements can be obtained on a wide spectrum of carbon-containing samples including charcoal, wood, marine and freshwater shell, bone and antler, peat and organic-bearing sediments, carbonate

deposits such as marl, tufa, and caliche, as well as carbonates and CO_2 dissolved in ocean, lake and groundwaters. With a half-life of approximately 5700 years, the ^{14}C method can be routinely employed in the age range of about 300 to between 40,000 and 50,000 years for samples in the range of 1–10 g of carbon using conventional decay or beta counting. With isotopic enrichment and larger sample sizes, ages up to 75,000 years have been measured. The use of accelerator mass spectrometry (AMS) for direct or ion counting of ^{14}C permits measurements to be obtained routinely on samples of several milligrams of carbon (with special efforts, on micrograms of carbon) and potentially may permit a significant extension of the ^{14}C time frame to as much as 100,000 years if stringent requirements for the exclusion of modern and *in situ* contamination can be met.

I. ELEMENTS OF THE RADIOCARBON (^{14}C) METHOD

A. Basic Principles

Carbon has three naturally occurring isotopes, two of which are stable (^{12}C, ^{13}C) and one (^{14}C) which is unstable or radioactive. The term radiocarbon is a contraction of the term "radioactive carbon." Radiocarbon decays by emitting a beta particle to form ^{14}N and a neutrino. The basis of the ^{14}C dating method can be outlined in terms of the production, distribution, and decay of ^{14}C (Fig. 1).

The natural production of ^{14}C is a secondary effect of cosmic-ray bombardment in the upper atmosphere. Following production, it is oxidized to form $^{14}CO_2$. In this form, ^{14}C is distributed throughout the earth's atmosphere. Most of it is absorbed in the oceans, while a small percentage becomes part of the terrestrial biosphere by means of the photosynthesis process and the distribution of carbon compounds through the different pathways of the earth's carbon cycle. Metabolic processes maintain the ^{14}C content of living organisms in equilibrium with atmospheric ^{14}C. However, once metabolic processes cease—as at the death of an animal or plant—the amount of ^{14}C will begin to decrease by decay at a rate measured by the ^{14}C half-life.

The *radiocarbon age* of a sample is based on measurement of its residual ^{14}C content. For a ^{14}C age to be equivalent to its actual or calendar age at a reasonable level of precision, a set of assumptions must hold within relatively

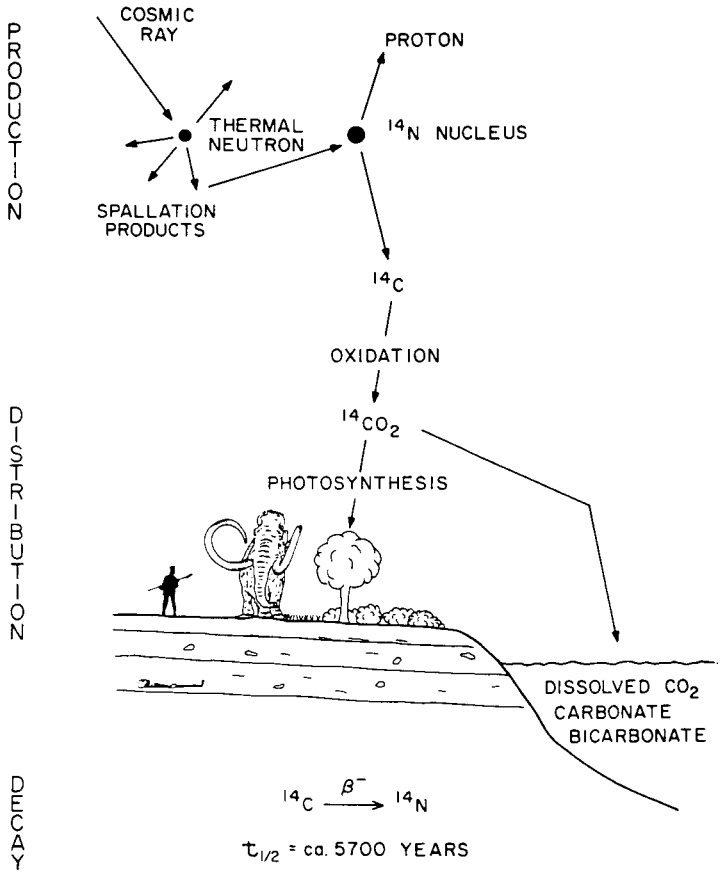


FIGURE 1 Radiocarbon dating model: production, distribution, and decay of ^{14}C. [From Taylor, R. E. (1987). "Radiocarbon Dating: An Archaeological Perspective," Academic Press, FL.]

narrow limits. These assumptions include: (i) the concentration of ^{14}C in each carbon reservoir has remained essentially constant over the ^{14}C time scale; (ii) there has been complete and rapid mixing of ^{14}C throughout the various carbon reservoirs on a worldwide basis; (iii) carbon isotope ratios in samples have not been altered except by ^{14}C decay since these sample materials ceased to be an active part of one of the carbon reservoirs—as at the death of an organism; (iv) the half-life of ^{14}C is accurately known, and, (v) natural levels of ^{14}C can be measured to appropriate levels of accuracy and precision.

The ^{14}C dating technique was developed at the University of Chicago in the period immediately following World War II by Willard F. Libby and his collaborators James R. Arnold and Ernest C. Anderson. Libby received the 1960 Nobel Prize in Chemistry for the development of the method.

B. Conventions

Radiocarbon age estimates are generally expressed in terms of a set of widely accepted parameters that define a *conventional radiocarbon age*. These parameters include (i) the use of 5568 (5570) years as the ^{14}C half-life even though the actual value is probably closer to 5730 years, (ii) the direct or indirect use of standard preparations of oxalic acid originally distributed by the United States National Bureau of Standards (now National Institutes of Standards and Technology) as a contemporary or modern reference standard to define a "zero" ^{14}C age, (iii) the use of A. D. 1950 as the zero point from which to count ^{14}C time, (iv) a correction or normalization of ^{14}C in all samples to a common $^{13}C/^{12}C$ value to account for fractionation effects, and (v) an assumption that ^{14}C in all reservoirs has remained constant over the ^{14}C time scale. In addition, a conventional understanding is that each ^{14}C determination should be accompanied by an expression that provides an estimate of the *experimental* or *analytical uncertainty*. Since statistical constraints associated with the measurement of ^{14}C is usually the dominant component of the experimental uncertainty, this value is sometimes informally referred to as the "statistical error" and is, by convention, expressed at the ± 1 sigma level. The experimental error should be suffixed to all appropriately documented ^{14}C age estimates.

For samples from some carbon reservoirs, the conventional contemporary standards may not define a zero ^{14}C age. A *reservoir corrected radiocarbon age* can sometimes be calculated by documenting the apparent age exhibited in control samples and correcting for the observed deviation. Reservoir effects are most often observed in samples from marine environments and fresh water lakes. A *calibrated radiocarbon age* takes into consideration the fact that ^{14}C activity in living organisms has not remained constant over the ^{14}C time scale because of secular variations effects.

II. MAJOR ANOMALIES

A. Secular Variation Effects

One of the most fundamental assumptions of the ^{14}C method is the requirement that natural ^{14}C concentrations in materials of "zero ^{14}C age" in a particular carbon reservoir are equivalent to that which has been characteristic of living organisms in that same reservoir over the entire ^{14}C time scale. Generally this assumption is seen to require an equilibrium or steady-state relationship in which the production and decay rates have been in approximate balance. Since the decay rate of ^{14}C is constant, the principal variables affecting equilibrium conditions would be changes in the atmospheric production rate of ^{14}C, long- and short-term climatic perturbations and effects related to variations in the parameters of the carbon cycle such as reservoir sizes and rates of transfer of ^{14}C between different carbon reservoirs.

Initially, tests of the validity of the assumption of constant ^{14}C concentration in living organics over time have focused on the analyses of the ^{14}C activity of a series of historically or dendrochronologically dated samples. The data base which initially provided the most valuable information on secular variation in the ^{14}C time series was ^{14}C determinations obtained on tree-ring dated wood, particularly that obtained on samples of the bristlecone pine (*Pinus longaeva*) from the western United States. A long series of paired ^{14}C/dendrochronologically dated samples of bristlecone and other species of wood documented the offset between "radiocarbon years" and "solar or calender years" indicating that ^{14}C ages were "too young" for most of the Holocene by as much as 1000 years when compared with the tree-ring derived ages.

More recently, comparisons of ^{14}C with dendrochronological data based on Irish oak and German oak and pine for the earliest portion and Douglas fir, sequoia and bristlecone pine for later periods document about 11,800 years of solar time. For the period before dendrochronological controls are currently available, paired uranium/thorium ($^{234}U/^{230}Th$) and ^{14}C samples from cores drilled into coral formations provide the principal data on which a late Pleistocene ^{14}C calibration curve has been extended to about 24,000 solar years BP or, as expressed in ^{14}C time, to about 20,300 BP. Figure 2 represents a plot of the off-set between ^{14}C and assumed "true" ages based primarily on combined dendrochronological/^{14}C and coral uranium-series/^{14}C data.

With the extension of the ^{14}C calibration framework using the uranium-series data on corals, it has now become apparent that what was thought of initially as a

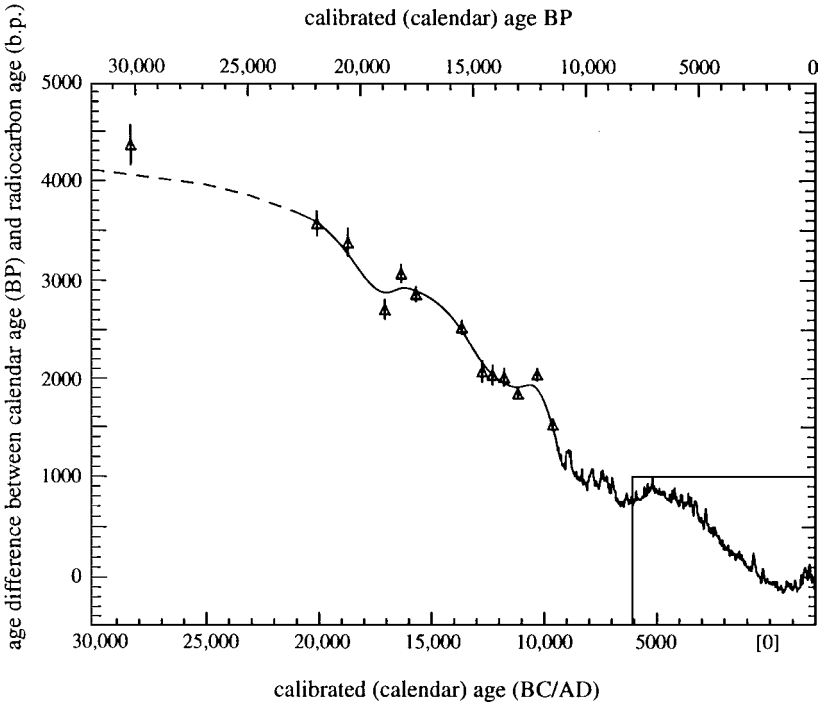


FIGURE 2 Characterization of Late Pleistocene and Holocene deviation of ^{14}C from dendrochronologial-, uranium-series-, and varve counted marine sediment-based age data. Box in lower right corner indicates original bristlecone pine-based indications of ^{14}C deviations from solar or "real" time. [From Taylor *et al.* (1996). "Development and extension of the calibration of the radiocarbon time scale," *Quaternary Sci. Rev.* (*Quaternary Geochronol.*) **15,** 655–668.]

"sine-wave"-like characterization of the variation in the ^{14}C time scale (illustrated in the lower right hand corner of Fig. 2) during the Holocene was an artifact of the limited time frame documented by the initial Holocene tree-ring/^{14}C data. Based on the expanded calibration data, it now appears that the long-term secular variation ^{14}C anomaly can be characterized as representing a slow-decay function on which has been superimposed middle- and short-term ("de Vries effects") perturbations.

A fuller rendering of the entire Holocene ^{14}C time scale now permits researchers to review more precisely the timing and characteristics of the de Vries "wiggles" over the last 10 millennia. Figure 3 plots the series of defined Holocene "time warps" reflecting the time ranges of the de Vries effect perturbations. Figure 3A represents 0 to 5000 BP and Fig. 3B represents the 5000 to 10,000 BP period in ^{14}C years.

In Fig. 3, twelve major and five intermediate de Vries effect perturbations can be identified. Major de Vries effects have been defined as warps exceeding 250 solar years; intermediate de Vries effects are those that exhibit ranges in excess of 140 solar years, and minor de Vries effects have ranges of less than 140 solar years. The 17 major and intermediate de Vries perturbations have been assigned Roman numeral and letter combinations. Roman numerals identify the ^{14}C millennium, i.e., I = 0 to 1000 BP, II = 1000 to 2000 BP, while lower case letters identify the perturbation in chronological order within each ^{14}C millennial period.

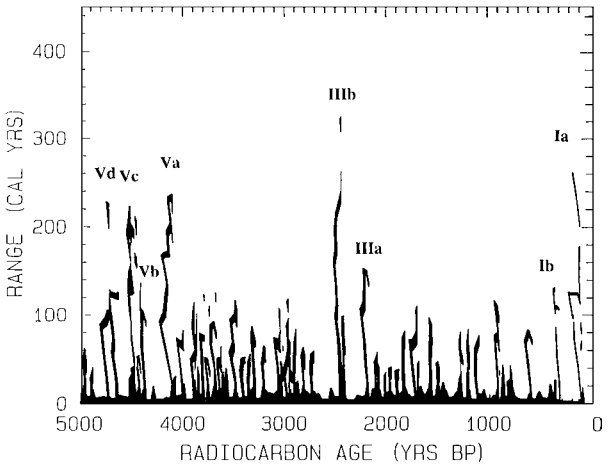


FIGURE 3A Holocene de Vries effects: Ranges ("warps") in calendar years obtained from the calibration of conventional ^{14}C ages: 0–5000 BP. [From Taylor *et al.* (1996). "Development and extension of the calibration of the radiocarbon time scale," *Quaternary Sci. Rev.* (*Quaternary Geochronol.*) **15,** 655–668.]

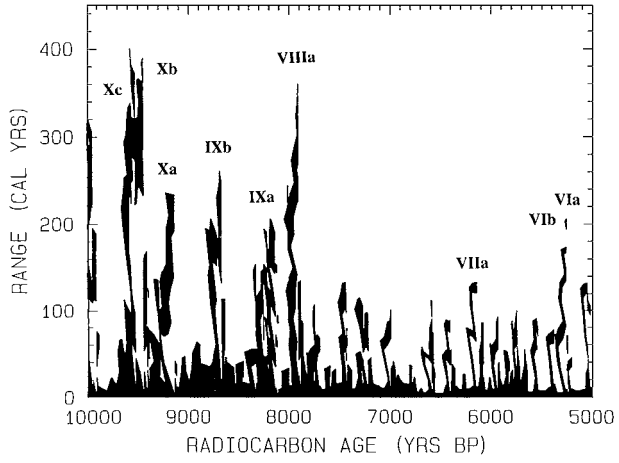


FIGURE 3B Holocene de Vries effects: Ranges ("warps") in calendar years obtained from the calibration of conventional ^{14}C ages: 5000–10,000 BP. [From Taylor *et al.* (1996). "Development and extension of the calibration of the radiocarbon time scale," *Quaternary Sci. Rev.* (*Quaternary Geochronol.*) **15,** 655–668.]

B. Reservoir, Fractionation, and Contamination Effects

One of the important features of the ^{14}C method is the potential of the worldwide comparability of ^{14}C values. For this potential to be realized, however, the initial ^{14}C activity in samples of identical age must be, within statistical limits, identical—each sample must begin with the same "zero" age ^{14}C activity. There are a number of situations where this condition is not met.

Reservoir effects occur when initial ^{14}C activities in samples of identical age but from different carbon reservoirs exhibit significantly different ^{14}C concentrations. In some cases, living samples from some environments exhibit apparent ^{14}C "ages" due to the fact that a significant percentage of the ^{14}C in these samples do not draw their carbon directly from the atmosphere. For example, in the case of samples from some fresh water lakes, the ^{14}C activity in living samples can be significantly reduced due to the dilution of the ^{14}C activity in lake waters with carbon from limestone in the lake beds which is Paleozoic in origin with an age in excess of several hundred million years. Carbon derived from these lake beds contained no measurable ^{14}C. This ^{14}C dilutes the atmospheric ^{14}C concentrations for materials obtaining their carbon primarily from the lake waters and gives these materials an apparent age.

In some marine environments, the ^{14}C activity in living shells is decreased as a result of dilution of ^{14}C upwelled from the deeper portions of the oceans where ^{14}C activities in bottom waters have been reduced because of relatively long residence times. Examples of other samples influenced by reservoir effects include wood and plant materials growing adjacent to active volcanic fumarole vents or where magmatic fossil CO_2 is being injected directly into lake waters from gas springs. Reservoir effects can range from a few hundred to a few thousand years depending upon specific circumstances.

For the ^{14}C method, the physical phenomena used to index time are changes in $^{14}C/^{12}C$ ratio. If carbon had only two naturally occurring isotopes only these would enter in calculations. However, as we have noted, carbon has three naturally occurring isotopes. Accurate estimates of the age of a sample using the ^{14}C method assumes that no change has occurred in the natural carbon isotope ratios except by the decay of ^{14}C. Several physical effects other than radioactive decay can alter the carbon isotope ratios in samples. The most commonly discussed problem involves *contamination* of samples in which carbon-containing compounds not indigenous to the original organic material are physically or chemically incorporated into a sample matrix. A number of sample pretreatment procedures have been developed by radiocarbon laboratories to deal with the problem of contamination. These procedures are sample specific and involve a wide range of chemical procedures designed to exclude non-*in situ* organics.

A related problem involves the *fractionation* of the carbon isotopes under natural conditions or because of procedures used in the laboratory during the preparation and measurement of samples. Fractionation involves alternation in the ratios of isotopic species as a function of their atomic mass. In the case of the carbon isotopes, ^{14}C has a mass about 15% greater than ^{12}C and thus ^{14}C is "heavier" than ^{12}C. During natural biochemical processes such as photosynthesis, the two "lighter" carbon isotopes are preferentially incorporated into sample materials. Because of this, variations in $^{14}C/^{12}C$ ratios can occur that have noting to do with the passage of time. Rather these variations are a function of which part of the carbon reservoir that samples are derived. By convention, fractionation effects are dealt with by measuring ^{13}C values in samples and normalizing ^{14}C values onto a common ^{13}C scale

C. Recent Anthropogenic ^{14}C Effects

Several factors contribute to make it very difficult to employ ^{14}C data to assign unambiguous calendar age estimates to materials dating to the last 300 years. First of all, de Vries effects are particularly pronounced during this period. In addition, during the 19th and 20th centuries, ^{14}C concentrations were seriously affected by human activities. As the result of the combustion of fossil fuels (Suess, Fossil Fuel, or Industrial Effect), ^{14}C "dead" CO_2 was released into the atmosphere, diluting the ^{14}C activity in biological materials by about 3%. In the post-World War II period, as a result of the detonation of thermonuclear devices in the atmosphere (Atomic Bomb or Nuclear

Effect) artificial or "bomb" ^{14}C was injected into the atmosphere. Radiocarbon concentrations in the atmosphere between about 1950 and 1965 almost doubled. A moratorium on testing in the atmosphere in 1963 halted testing in the atmosphere by most of the nuclear powers. This has allowed ^{14}C to begin the process of reestablishing a new atmospheric ^{14}C equilibrium.

Due to the combination of these two effects, a ^{14}C age estimate may be reported as simply "modern" when the conventional ^{14}C value is less than 150 years. Because of the post-16th-century de Vries and Suess effects, the ^{14}C method can not, except in rare instances, be used to assign specific age to materials living between 1650 and 1950 other than that they were limiting during this 300-year period. However, the presence of "bomb" ^{14}C in materials can be used to determine that these materials were living after 1950.

III. MEASUREMENT TECHNIQUES

A. Gas or Liquid Scintillation Decay or Beta Counting

The challenge of accurately and precisely measuring natural ^{14}C concentrations in organic samples proceeds from the consequences of several factors. The most significant is that natural ^{14}C occurs in extremely low concentrations. The natural isotopic composition of modern carbon is about 98.9 ^{12}C, 1.1% ^{13}C, and 10^{-10}% ^{14}C. One naturally produced ^{14}C atom exists for about every 10^{12} (1,000,000,000,000) ^{12}C atoms in living materials. This concentration decreases by a factor of 2 for every 5700-year period following the death of an organism.

All of the pioneering research of Libby and his coworkers and essentially all ^{14}C age estimates obtained up until the early 1980s were obtained using decay counting methods. Decay counting involves the use of either elemental or solid carbon (which became obsolete in mid 1950s), gas (CO_2, CH^4, C_2H_2, and rarely C_2H_6) or liquid scintillation counting. Radiocarbon decays through weak beta decay. Solid carbon and gas decay counting employs a type of Geiger–Müller (GM) instrument to detect the decay of ^{14}C in a sample. Because of the relatively low specific activity of natural ^{14}C and the relatively high environmental count rate (primarily due to cosmic radiation), the most important technical development that made natural ^{14}C measurements practical involved the development by Libby's group of an anticoincidence circuit linking together a GM instrument containing the sample with a set of surrounding ring of GM counters all placed together within a physical shield of lead or iron. By comparing the time of arrival of a pulse from the sample detector with that from the surrounding ring, decay events could be distinguished from events caused by environmental ionizing radiation.

Liquid scintillation counting takes advantage of the fact that in the presence of ionizing radiation, certain types of organic compounds (scintillators) emit short burst of light energy which can be converted into an electrical current in a photomultiplier tube. In liquid scintillation applications in the measurement of natural ^{14}C, samples are converted to benzene and the benzene is introduced into a sample cell along with scintillator and other chemicals to increase the sensitivity of photomultiplier detection. Most gas and liquid scintillation ^{14}C systems typically require the sample sizes in the range from 1 to 20 g of carbon. Measurement of samples takes from several days to several weeks. Several laboratories have constructed mini- and micro-size gas counters which can accommodate sub-gram-size samples but counting times for such systems are typically measured in months.

B. Accelerator Mass Spectrometry Direct or Ion Counting

In conventional decay counting systems, a very small fraction of the ^{14}C atoms present in a sample decay and are measured. Since the late 1960s, it was recognized that significantly higher efficiencies of atom-by-atom detection such as that employed in mass spectrometers would allow the use of sample sizes several orders of magnitude below that which is typically possible with decay counting (measured in milligrams rather than grams of carbon), would permit much shorter counting times, and could provide a potential method of extending the maximum ^{14}C age range beyond that possible with decay-counting techniques. Efforts to achieve direct or ion counting of ^{14}C with conventional mass spectrometers were frustrated because of high backgrounds until the late 1970s, when experiments with particle accelerators were initiated.

Two types of particle accelerators have been employed for accelerator mass spectrometry (AMS) ^{14}C measurements: cyclotrons and tandem electrostatic accelerators. Cyclotrons, which sends particles along a spiral trajectory, successfully detected ^{14}C and made initial measurements. However, consistent results proved difficult to achieve despite years of effort. Routine AMS-based natural level ^{14}C measurements were achieved using tandem accelerators. Use of negative ions discriminated against ^{14}N since nitrogen does not form negative ions that live long enough to pass through the accelerator to the detector. Thus, an important source of background was essentially eliminated. Also, the process of ion acceleration in a tandem accelerator breaks up mass 14 molecules which would otherwise interfere with ^{14}C detection. A schematic

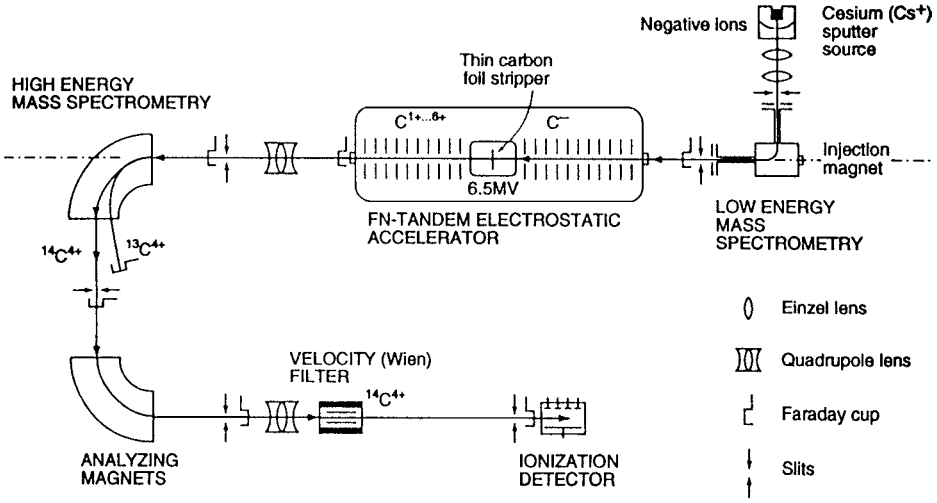


FIGURE 4 Schematic of tandem accelerator used for direct accelerator mass spectrometry ^{14}C measurements at the Center for Accelerator Mass Spectrometry, University of California, Lawrence Livermore National Laboratory. [From Taylor, R. E. (1997) Radiocarbon dating. *In* (R. E. Taylor and M. Aitken, eds.), "Chronmetric Dating in Archaeology," Plenum Press, New York.]

of the elements of a tandem AMS system is illustrated in Fig. 4. While the ultimate ^{14}C maximum dating range using AMS technology may approach 100,000 years, current requirements for most AMS ^{14}C systems that samples be converted to graphitic carbon introduces sufficient modern contamination so that the current maximum ages that can be measured on AMS systems with conventional samples range between 40,000 and 60,000 years.

SEE ALSO THE FOLLOWING ARTICLES

CARBON CYCLE • GEOLOGIC TIME • PALYNOLOGY • RADIOACTIVITY • RADIOMETRIC DATING • TIME AND FREQUENCY

BIBLIOGRAPHY

Bard, E. (1998). Geochemical and geophysical implications of the radiocarbon calibration. *Geochim. Cosmochim. Acta* **62,** 2025–2038.

Geyh, M. A., and Schleicher, H. (1990). Radiocarbon Dating. *In* "Absolute Age Determination: Physical and Chemical Dating Methods and their Application," pp. 162–180, Springer-Verlag, Berlin.

Gove, H. (1992). The history of AMS, its advantages over decay counting: applications and prospects. *In* "Radiocarbon After Four Decades: An Interdisciplinary Perspective" (R. E. Taylor, R. Kra, and A. Long, A., eds.), Springer-Verlag, New York.

Hedges, R. E. M., and Gowlett, J. A. J. (1986). Radiocarbon dating by accelerator mass spectrometry. *Scientific American* **254,** 100–107.

Mook, W. G., and van der Plicht, J. (eds.) (1998). Proceedings of the 16th International Radiocarbon Conference. *Radiocarbon* **40,** 1–1040.

Polach, D. (1988). "Radiocarbon Dating Literature, the First 21 Years, 1947–1968," Academic Press, London.

Stuiver, M., von der Plicht, J., and Long, A. (eds.) (1998). Calibration Issue INTCAL 98. *Radiocarbon* **40,** 1041–1160.

Taylor, R. E. (1987). "Radiocarbon Dating: An Archaeological Perspective," Academic Press, Orlando.

Taylor, R. E. (1997). Radiocarbon dating. *In* "Chronometric Dating in Archaeology" (R. E. Taylor and M. J. Aitken, eds.), Plenum Press, New York.

Taylor, R. E. (2000). Fifty Years of Radiocarbon Dating. *American Scientist* **88,** 60–67.

Taylor, R. E., Stuiver, M., and Reimer, P. J. (1996). Development and extension of the calibration of the radiocarbon time scale: Archaeological applications. *Quaternary Sci. Rev.* (*Quaternary Geochronol*) **15,** 655–668.

Taylor, R. E., Kra, R., and Long, A. (eds.) (1992). "Radiocarbon After Four Decades: An Interdisciplinary Perspective," Springer-Verlag, New York.

Radiometric Dating

Marvin Lanphere
U.S. Geological Survey

GLOSSARY

Concordia The locus of all concordant U–Pb ages on a $^{206}Pb/^{238}U$ vs $^{207}Pb/^{235}U$ diagram.
Decay constant Probability that a radioactive atom will decay in unit time.
Half-life Time required for half of a given quantity of radioactive atoms to decay.
Isochron A line on a parent–daughter diagram defined by all points having the same age.
Isotope dilution Analytical method used to measure amounts of radioactive parent isotope and radiogenic daughter isotope in a sample.
Isotopes Atoms that have the same number of protons in the nucleus but different numbers of neutrons.
Radioactive decay Process in which an unstable nucleus is transformed into an isotope of another element.

RADIOMETRIC DATING comprises all methods of age determination of rocks and minerals based on nuclear decay of naturally occurring radioactive isotopes. The 84 elements found in nature are represented by 269 stable isotopes and 70 radioactive isotopes. Eighteen of the radioactive isotopes have long half-lives and have survived since the elements of the Solar System were manufactured. These long-lived radioactive isotopes are the basis for radiometric dating. Most of the modern dating methods depend on the decay of a radioactive parent to a stable daughter product. The remaining 52 radioactive isotopes, including ^{14}C which is the basis of an important dating method, have short half-lives but are either continuously created by nuclear reactions or are produced by decay of long-lived radioactive isotopes.

I. HISTORY OF DEVELOPMENT OF RADIOMETRIC DATING METHODS

The initial event that led to the development of radiometric dating methods was the discovery, by Wilhelm Roentgen in 1895, of X-rays, which are produced when a beam of electrons strikes a solid target. It was thought that the emission of X-rays might be related to the phosphorescence produced by cathode rays on the walls of a vacuum tube. In 1896 Henri Becqueral, while studying phosphorescence, discovered that uranium salts emitted radiation that had properties similar to X-rays. Two years later Marie and Pierre Curie discovered that thorium also emitted radiation, and they named the new phenomenon

"radioactivity." The Curies also determined that radioactivity was a property of atoms, not molecules, and depended only on the presence of the radioactive element.

In 1902 Ernest Rutherford and Frederick Soddy of McGill University formulated a general theory in which they proposed that the atoms of radioactive elements were unstable and decayed spontaneously to other elements accompanied by the emission of alpha or beta particles. In 1905 Rutherford further suggested that radioactivity could possibly be used as a geological timekeeper if the production rate of helium from known weights of radioactive elements were known experimentally.

By 1902, many different radioactive substances had been discovered, and, for example, it was obvious that the radioactivity of uranium (U) and thorium (Th) was a composite effect produced by several different radioactive isotopes in a cascade of decays. In 1905 the American chemist, Bertram Boltwood, suggested that lead (Pb) was the final (nonradioactive) product of the radioactive decay of uranium. Boltwood then proceeded to show in 1907 that in minerals of the same age the Pb/U ratio is constant whereas among minerals of different ages the Pb/U ratio is different; the older the mineral the greater the ratio. Boltwood then measured a number of "chemical ages" based on the amounts of U and Pb in several minerals. The rate of disintegration of U was not known, but Rutherford had estimated the disintegration rate of radium (Ra), which is a decay product of U. If the decay of U to Ra to Pb was in secular equilibrium, then ages could be calculated using the estimated disintegration rate of Ra and the Ra/U ratio at equilibrium. These ages had inherent uncertainties because *isotopes*, atoms with different atomic weights but identical atomic numbers, had not been discovered. Isotopes of an element have the same number of protons in the nucleus and behave similarly chemically but have different numbers of neutrons in the nucleus. It also was not known in 1907 that lead is also a product of the radioactive decay of Th.

II. PRINCIPLES OF RADIOMETRIC DATING

Atoms of elements are composed of protons, neutrons, and electrons. Only certain combinations of protons and neutrons yield stable nuclides; all other combinations are unstable. *Radioactive decay* is the process in which an unstable nucleus either ejects or captures particles transforming the radioactive nuclide into another element. In some cases the daughter nuclide produced by radioactive decay also is unstable and radioactive decay continues through as many steps as necessary to produce a stable nuclide.

Only three types of radioactive decay are important in radiometric dating. The most common type is beta (β^-) decay in which a beta particle, which is an energetic electron, is ejected from the nucleus. The β^- particle is created by the breakup of a neutron into a proton, the β^- particle, and an antineutrino. The β^- particle and antineutrino are ejected from the nucleus and the proton remains in the nucleus. The number of protons in the nucleus is increased by one, the number of neutrons is decreased by one, and the atomic mass is unchanged.

A second type of decay is electron capture or ec. In ec decay an electron from the innermost electron shell falls into the nucleus converting a proton into a neutron. The number of neutrons in the nucleus is increased by one, the number of protons is decreased by one, and the atomic mass is unchanged.

The third type of decay is alpha (α) decay, which primarily occurs among heavy elements. In α decay the nucleus ejects an α particle, a particle consisting of two protons and two neutrons. Thus, the number of protons and number of neutrons are each decreased by two and the atomic mass is decreased by four.

Radioactive decay is a statistical process in which the number of atoms that disintegrate per unit time, $-d\mathrm{N}/dt$, is proportional to the number of atoms originally present, N. Thus,

$$-d\mathrm{N}/dt = \lambda \mathrm{N}, \tag{1}$$

where λ, the *decay constant* represents the probability that an atom will decay in unit time. This equation can be integrated to yield

$$\mathrm{N} = \mathrm{N}_0\, \mathrm{e}^{-\lambda t}, \tag{2}$$

which is the basic radioactive decay formula; N_0 is the initial number of atoms present and N is the number of atoms at time *t*. A radioactive nuclide is characterized by the rate at which it disintegrates. For long-lived nuclides used for radiometric dating the decay constant and the half-life are used as a measure of this rate. The decay constant was introduced above. The *half-life*, $t_{1/2}$, is the time required for half of a given quantity of radioactive atoms to decay.

Equation (2) is not very useful for radiometric dating because there is no way to determine the initial number of parent atoms, N_0. There is an easy solution to this problem, however, because the sum of the parent atoms left, P_t, and the daughter atoms formed, D_t, must always equal the original number of parent atoms, N_0:

$$\mathrm{N}_0 = \mathrm{P}_t + \mathrm{D}_t. \tag{3}$$

If we now substitute $\mathrm{P}_t + \mathrm{D}_t$ for N_0 in equation 2 where $\mathrm{P}_t = \mathrm{N}$, then

$$P_t = (P_t + D_t)e^{-\lambda t} \quad (4)$$

and

$$D_t = P_t(e^{-\lambda t} + 1) \quad (5)$$

This equation can be solved for t, giving

$$t = 1/\lambda \log_e(D_t/P_t + 1) \quad (6)$$

which is the basic radiometric-age equation and contains only quantities that can be measured today in the laboratory.

If a rock or mineral to be dated contained none of the daughter isotope at the time of formation, then the age of the rock or mineral could be determined at some later time using Eq. (6) after D_t and P_t have been measured. However, if the rock or mineral incorporated some of the daughter isotope during formation, then this initial amount of the daughter isotope, D_0, would have to be subtracted from the total measured amount in order to yield the correct age:

$$t = 1/\lambda \log_e [(D_t - D_0)/P_t + 1], \quad (7)$$

where D_t is now the total number of daughter atoms present.

Several important assumptions are made when Eq. (7) is used to calculate the age of a rock or mineral. These are:

1. The decay of a radioactive parent isotope takes place at a constant rate regardless of its chemical or physical environment. Radioactive decay is a nuclear phenomenon and should be independent of environmental conditions such as temperature and pressure. It has been found that radioactive decay is constant over wide ranges of temperature and pressure.
2. The present day abundance of the radioactive parent isotope to the total parent element is the same in all materials. Because the radioactive parent is decaying the abundance ratio has, of course, decreased through time. However, as shown in Eq. (5) it is only necessary to know the amount of parent isotope present in a rock or mineral today.
3. The radiogenic daughter isotope measured in a sample was produced by *in situ* decay of the radioactive parent isotope in the interval since the rock crystallized or was recrystallized. Violations of this assumption are not uncommon and must be tested for each dating method.
4. Corrections can be made for nonradiogenic daughter isotopes in the rock or mineral being dated. Various ways of evaluating this assumption are available including the use of isotope correlation diagrams described below.
5. The rock or mineral has been a closed system since t_0, the time that the rock or mineral formed. That is, there has been no gain of radioactive parent isotope or radiogenic daughter isotope except for that which results from the radioactive decay of the parent isotope. The validity of this assumption depends strongly on geologic conditions and each case must be decided separately.

III. RADIOMETRIC DATING METHODS

Radiometric dating relies on the accumulation with time of daughter isotopes created by the radioactive decay of a certain amount of parent in a closed system. This relationship is expressed by either Eqs. (6) or (7) depending on the absence or presence of an initial quantity of daughter isotope. In practice, the presence of initial daughter isotope generally precludes the use of simple accumulation methods.

Radiometric dating depends on the quantitative measurement of the amounts of radioactive parent isotope and radiogenic daughter isotope in a sample. These amounts are often very small. A very sensitive analytical method used to make these measurements is called *isotope dilution*. A known quantity of a tracer, which has an isotopic composition different from that of the natural element, is mixed with the sample being analyzed. For example, the $^{40}Ar/^{38}Ar$ ratio in air is 1581, and tracer ^{38}Ar produced by gaseous diffusion of atmospheric Ar has a $^{38}Ar/^{40}Ar$ ratio of 100,000. The isotopic composition of the mixture is measured on a mass spectrometer, and the amount of the natural element in the sample can be calculated. A simple analogy to isotope dilution can be drawn using the classic standbys, apples and oranges. Suppose one had an unknown number of oranges and wanted to find how many oranges there were. One could mix a known number of apples with the oranges, take a representative sample of the mixture, determine the ratio of oranges to apples in the representative mixture, and then easily calculate the number of oranges. Tracers are produced artificially. Tracers for solid elements commonly are produced by isotope separation in large mass spectrometers called calutrons. Isotope dilution can be applied to any element that has more than one isotope.

A. The K–Ar Method

The K–Ar method, which is based on the decay of ^{40}K to ^{40}Ar, probably has been the most widely used radiometric dating technique available to geologists. Potassium is the sixth most abundant element in the earth's crust, and most rocks and minerals contain sufficient potassium for

TABLE I Principal Parent and Daughter Isotopes Used to Determine the Ages of Rocks and Minerals

Parent isotope (radioactive)	Daughter isotope (stable)	Half-life (Ma)	Decay constant ($Year^{-1}$)
^{40}K	$^{40}Ar^{a}$	1,250	5.81×10^{-11}
^{87}Rb	^{87}Sr	48,800	1.42×10^{-11}
^{147}Sm	^{143}Nd	106,000	6.54×10^{-12}
^{176}Lu	^{176}Hf	35,900	1.93×10^{-11}
^{187}Re	^{187}Os	43,000	1.612×10^{-11}
^{232}Th	^{208}Pb	14,000	4.948×10^{-11}
^{235}U	^{207}Pb	704	9.8485×10^{-10}
^{238}U	^{206}Pb	4,470	1.55125×10^{-10}

[a] ^{40}K also decays to ^{40}Ca, for which the decay constant is 4.962×10^{-11} $year^{-1}$, but that decay is not used for dating. The half-life is for the parent isotope and includes both decays.

age measurements. Potassium has one radioactive isotope, ^{40}K, and two stable isotopes, ^{39}K and ^{41}K. The value of the half-life of ^{40}K, 1.25 Ga (10^9 years) (Table I), also is favorable so that rocks both as old as the Earth, 4.5 Ga, and as young as a few thousand years are amenable to dating using the K–Ar method. The radioactivity of ^{40}K is interesting in that this parent isotope undergoes decay by two modes: by electron capture to ^{40}Ar and by β^- emission to ^{40}Ca. The ratio of ^{40}K atoms that decay to ^{40}Ar to those that decay to ^{40}Ca is 0.117, which is called the branching ratio. ^{40}Ca is the most abundant isotope of Ca and because Ca is also very abundant in rocks and minerals, it generally is not possible to determine the amount of ^{40}Ca present initially. Thus, the ^{40}K–^{40}Ca dating method is rarely used.

The K–Ar method is the only decay scheme that can be used without major concern about the initial presence of daughter isotope. ^{40}Ar is an inert gas that does not combine chemically with any other element and it escapes easily from rocks when they are heated. For example, when a rock is molten in a magma chamber ^{40}Ar produced by decay of ^{40}K escapes from the liquid. After a rock has crystallized from the liquid and has cooled, radiogenic ^{40}Ar accumulates in crystal lattices of minerals over time. If the rock is heated or melted at some later time, then some or all of the ^{40}Ar may escape and the K–Ar clock will be partly or entirely reset.

Newly crystallized rocks rarely contain measurable initial ^{40}Ar so Eq. (6) applies to the K–Ar method. But, Eq. (6) must be modified to take the branching decay of ^{40}K into account.

$$t = 1/(\lambda_{ec} + \lambda_{\beta^-}) \log_e \left\{ {}^{40}Ar/{}^{40}K[(\lambda_{ec} + \lambda_{\beta^-})/\lambda_{ec})] + 1 \right\}, \quad (8)$$

where λ_{ec} and λ_{β^-} are the decay constants of the ec and β^- decays, respectively. If values for the decay constants are now substituted in Eq. (8), the K–Ar age equation becomes

$$t = 1.804 \times 10^9 \log_e \{9.543\,{}^{40}Ar/{}^{40}K + 1\}, \quad (9)$$

where t is in years.

Argon has three isotopes, ^{36}Ar, ^{38}Ar, and ^{40}Ar. Argon makes up nearly one percent of the earth's atmosphere, and can be incorporated in rocks, particularly extrusive volcanic rocks. A correction must be made for this atmospheric Ar in a sample and also in the experimental apparatus used to extract Ar from rocks and minerals. This correction is easily made by measuring the amount of ^{36}Ar in the experiment, multiplying by 295.5 which is the $^{40}Ar/^{36}Ar$ ratio in the atmosphere, and subtracting this result from the total amount of ^{40}Ar. What is left is the amount of radiogenic ^{40}Ar produced from decay of ^{40}K.

The K–Ar method works best on intrusive igneous and extrusive volcanic rocks that have crystallized from melts and have not been reheated. The method does not work well on metamorphic rocks because these rocks have been produced from other rocks by heat and pressure without being completely melted.

B. The $^{40}Ar/^{39}Ar$ Method

The $^{40}Ar/^{39}Ar$ dating method is a form of K–Ar dating in which a sample is irradiated in a nuclear reactor with fast neutrons to convert a fraction of the ^{39}K, which is the most abundant isotope of K, to ^{39}Ar. The reaction of a fast neutron, generally those with energies greater than 0.02 MeV, with a ^{39}K nucleus results in the addition of a neutron and the ejection of a proton, which changes the ^{39}K to ^{39}Ar. In the conventional K–Ar method the amounts of K and Ar in a sample must be determined quantitatively in separate experiments. In the $^{40}Ar/^{39}Ar$ method the ratio of radioactive parent to radiogenic daughter is determined by measuring the ratio of ^{40}Ar to ^{39}Ar in one experiment. Corrections must be made for atmospheric Ar and for interfering Ar isotopes produced by neutron reactions with Ca and other K isotopes.

An age can be calculated from $^{40}Ar/^{39}Ar$ data using an equation that is similar to Eq. (9):

$$t = 1.804 \times 10^9 \log_e (J\,{}^{40}Ar/{}^{39}Ar + 1), \quad (10)$$

where J is a constant that includes a factor for the fraction of ^{39}K converted to ^{39}Ar during irradiation. J is determined for each irradiation by irradiating a sample of known age as measured by the K–Ar method, which is called a flux monitor, together with the unknown sample and using Eq. (10) to calculate J for the monitor. This value for J applies to the unknown sample because both the monitor and unknown sample received the same neutron dose.

If an irradiated sample is completely melted, the Ar in the sample gives an age that is comparable to a

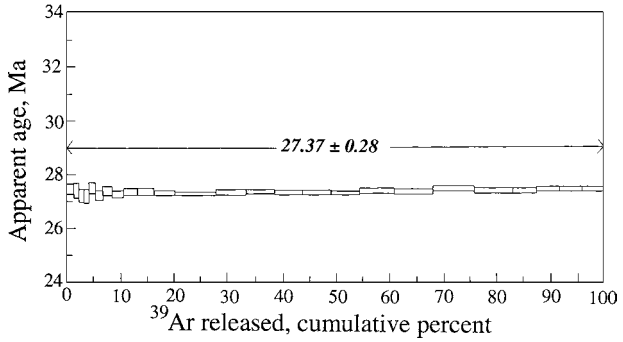


FIGURE 1 $^{40}Ar/^{39}Ar$ age spectrum. Horizontal line is apparent age for each temperature increment. Half of vertical dimension of increment boxes is estimated standard deviation of increment age. Error of weighted mean plateau age is a weighted standard deviation.

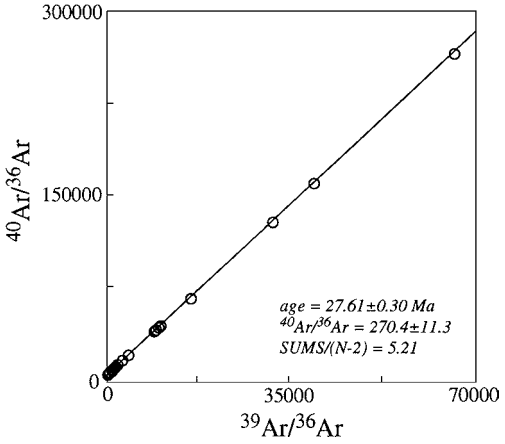


FIGURE 2 $^{40}Ar/^{39}Ar$ isochron diagram. Increment data shown as circles were used in the isochron fit; increment data shown as crosses were not used in the fit.

conventional K–Ar age. However, the precision of this total-fusion $^{40}Ar/^{39}Ar$ age generally is better than the precision of a conventional K–Ar age. The $^{40}Ar/^{39}Ar$ method can also be applied to smaller samples. But, the primary advantage of the $^{40}Ar/^{39}Ar$ method is that the sample can be heated to progressively higher temperatures and the Ar released at each temperature can be collected and analyzed separately. An age can then be calculated for each temperature increment. The series of ages from an incremental-heating experiment are often plotted versus the percentage of ^{39}Ar released. This type of diagram is called an age spectrum diagram. For an undisturbed sample the calculated ages are all the same and the age spectrum is a horizontal line. Real samples often are not completely ideal, and the age spectrum is more complex. In this case a series of contiguous gas samples can be selected that have the same age within experimental error and form a horizontal age spectrum that is called a plateau (Fig. 1). The experimental data can also be plotted on an $^{40}Ar/^{39}Ar$ *isochron* or isotope correlation diagram (Fig. 2). The data should fall on or near a straight line whose slope is equal to the ratio $^{40}Ar/^{39}Ar$ in Eq. (10) and whose intercept is the $^{40}Ar/^{36}Ar$ ratio of nonradiogenic Ar.

C. The Rb–Sr Method

The Rubidium–Strontium (Rb–Sr) method is based on the radioactivity of ^{87}Rb which undergoes β^- decay to ^{87}Sr with a half-life of 48.8 Ga (Table I). Rb is never a major constituent in minerals, but the chemistry of Rb is similar to that of K and Na that do form many common minerals. Thus, Rb occurs as a trace element in most rocks. Sr also occurs primarily as a trace element. Its chemistry is similar to that of Ca so Sr occurs in a wide variety of rocks.

Sr is present as a trace element in most minerals when they form. This contrasts with Ar which escapes easily from most molten rocks. Accurate accumulation ages can be calculated from Eq. (7) only for rare minerals that contain large amounts of Rb and a negligible amount of initial Sr. However, most rocks contain measurable amounts of initial Sr. An equation like (7) is of little use because the amount of initial ^{87}Sr cannot be determined. Rb–Sr dating is done using a graphical method, commonly called the isochron method, that eliminates the problem of initial Sr.

For a system closed to gain or loss of strontium and rubidium the age equation may be written in the form:

$$^{87}Sr_t = {}^{87}Sr_0 + {}^{87}Rb_t\,[e^{\lambda t} - 1], \tag{11}$$

and using the stable isotope ^{86}Sr as a convenient index isotope:

$$^{87}Sr_t/{}^{86}Sr_t = {}^{87}Sr_0/{}^{86}Sr_0 + {}^{87}Rb_t/{}^{86}Sr_t\,[e^{\lambda t} - 1]. \tag{12}$$

This equation shows that all closed systems of "age" t, which had a given initial $^{87}Sr_0/^{86}Sr_0$ ratio, will plot on a straight line (an isochron) in a $^{87}Sr/^{86}Sr$ vs. $^{87}Rb/^{86}Sr$ diagram which is known as the strontium evolution diagram. The intercept of this line for $^{87}Rb/^{86}Sr = 0$ gives the initial ratio $(^{87}Sr/^{86}Sr)_0$ which was common to those systems. The slope of the line is $[e^{\lambda t} - 1]$. All possible systems of the same age, but of differing initial $^{87}Sr/^{86}Sr$ ratios, will lie on parallel straight lines. The time t is the time since the various subsystems defining a particular isochron had the same $^{87}Sr/^{86}Sr$ value.

The time evolution of the various minerals in a rock in terms of the strontium evolution diagram is shown in

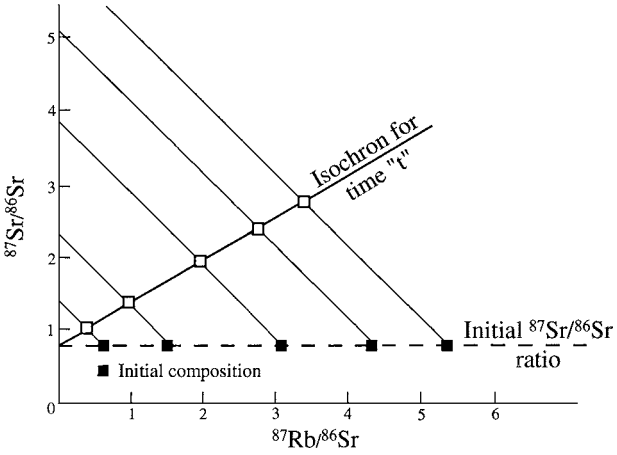


FIGURE 3 Strontium evolution diagram showing time evolution of $^{87}Sr/^{86}Sr$ and $^{87}Rb/^{86}Sr$ in mineral phases. Filled squares are minerals in a rock that have the same $^{87}Sr/^{86}Sr$ ratios but different $^{87}Rb/^{86}Sr$ ratios as the time the rock initially crystallized. As ^{87}Rb decays to ^{87}Sr isotopic compositions move along lines toward open boxes. A line through the open boxes is an isochron that gives the age of the rock at some time "t" after initial crystallization.

Fig. 3. If the minerals are closed to strontium and rubidium and all have the same initial strontium isotopic composition, but different Rb/Sr ratios, they will all initially lie on a horizontal straight line. As ^{87}Rb decays to ^{87}Sr each mineral will move along a straight line trajectory of slope:

$$d(^{87}Sr/^{86}Sr)/d(^{87}Rb/^{86}Sr) = -1 \text{ for all times.} \quad (13)$$

At a time t years subsequent to the condition of a homogeneous strontium isotopic composition for all phases the minerals will lie on an isochron as shown in Fig. 3.

This construction also applies to a series of cogenetic rocks which initially had the same strontium isotopic composition. In this case it is not necessary to assume that each individual mineral species is closed but that the rock samples analyzed represent closed systems. A closed system is a rock that has had no addition of radioactive parent isotope or loss of radiogenic daughter isotope. If a suite of rocks or minerals satisfies the linear relationship, this provides rather strong evidence that the systems analyzed had the same initial strontium isotopic composition and were closed to gain or loss of strontium and rubidium.

^{87}Rb has a long half-life, which means that the Rb–Sr dating method can be applied only to those rocks which are old enough so that measurable amounts of radiogenic ^{87}Sr have accumulated. The method is used mostly on rocks older than the Tertiary period, a few tens of million years. The Rb–Sr dating method is very useful on rocks that have had a complex history because if a rock is recrystallized during a metamorphic event, the isochron will give the age of the metamorphic event.

D. The Sm–Nd, Lu–Hf, and Re–Os Methods

The decays of ^{147}Sm (Samarium) to ^{143}Nd (Neodymium), ^{176}Lu (Lutetium) to ^{176}Hf (Hafnium), and ^{187}Re (Rhenium) to ^{187}Os (Osmium) all have long parental half-lives and very low natural abundances of the parent isotopes which means that the radiogenic daughter isotopes accumulate very slowly. Before about 1980 these decay schemes were of little value as radiometric dating methods. However, improvements in analytical techniques have permitted geochronologists to apply these three dating methods to certain types of problems.

The Sm–Nd method is the most widely used of these three methods. ^{147}Sm decays by α emission to ^{143}Nd with a half-life of 106 Ga (Table I). Sm and Nd are rare earths or lanthanide elements that have similar chemical behavior; natural geochemical processes do not produce separation of the elements. Thus, significant variations in the concentrations of Sm and Nd are not common. This means that significant quantities of initial Nd are present in all samples. Equation (7) is of no use and the isochron method must be used to extract age information. The rare earth elements occur as trace elements in most rocks and minerals although their concentrations are generally only a few parts per million.

Age determinations by the Sm–Nd method generally are made by analyzing separated minerals from a rock or cogenetic rocks whose Sm/Nd ratios vary sufficiently to define the slope of an isochron. The plot of $^{143}Nd/^{144}Nd$ vs $^{147}Sm/^{144}Nd$ is analogous to the strontium evolution diagram of the Rb–Sr dating method described above. The Sm–Nd method is well suited to dating mafic igneous rocks whereas the Rb–Sr method is more suitable to dating felsic igneous rocks that are enriched in rubidium and depleted in strontium. In addition, the rare earth elements are less mobile than the alkali metals and alkaline earths during regional metamorphism and chemical alteration. Thus, the Sm–Nd method can be used to reliably date rocks even though the rocks may have gained or lost Rb or Sr.

Lu, like Sm and Nd, is a rare earth element, but its radiogenic daughter is not. The Lu–Hf dating method is based on the β-decay of ^{176}Lu to ^{176}Hf with a half-life of 35 Ga (Table I). Lu occurs as a trace element in most rocks but in concentrations that rarely exceed 0.5 ppm. The common rock-forming minerals plagioclase, amphibole, pyroxene, and olivine have Lu concentrations of fractions of one part per million. Some accessory minerals such as apatite, garnet, and monazite contain several ppm of Lu and may be suitable for Lu–Hf dating. Zircon contains on average 24 ppm Lu; however, zircon is not suitable for dating because of its high Hf content. Minerals high enough in Lu and low enough in Hf that the Lu–Hf dating

method can be applied as a simple accumulation clock are rare, and the isochron method generally must be used.

The Re–Os dating method is based on the β-decay of ^{187}Re to ^{187}Os with a half-life of 43 Ga (Table I). Re and Os are metals whose abundance in igneous rocks is only about 0.5 parts per billion, which makes the method of little use in dating common rocks. However, Re is chemically similar to Mo (molybdenum) and like Mo is concentrated in the mineral molybdenite in areas of copper mineralization. The Re–Os method has been used to date some ore deposits. The method has also been used to date the metallic phases of some meteorites.

E. The U, Th, Pb Methods

The decay of uranium (U) and thorium (Th) to stable isotopes of lead (Pb) is the basis for several important dating methods. Uranium has three naturally occurring isotopes, ^{238}U, ^{235}U, and ^{234}U, all of which are radioactive. Thorium exists primarily as one radioactive isotope: ^{232}Th. Thorium also has five short-lived isotopes that are intermediate daughter products in the decays of ^{238}U, ^{235}U, and ^{232}Th. Each atom of ^{238}U that decays ultimately produces an atom of ^{206}Pb by emission of eight alpha particles and six beta particles. Each atom of ^{235}U that decays ultimately produces an atom of ^{207}Pb by emission of seven alpha particles and four beta particles. Each atom of ^{232}Th that decays ultimately produces an atom of ^{208}Pb by emission of six alpha particles and four beta particles. In these three decay series 43 isotopes of 12 elements are formed as intermediate daughters. It turns out that these intermediate daughters can be ignored insofar as the dating methods are concerned for two reasons. First, none of the intermediate daughter isotopes occurs in more than one series. This means that each decay chain always leads to the formation of a specific isotope of Pb. Second, the half-lives of ^{238}U, ^{235}U, and ^{232}Th are all much longer than the half-lives of their daughter products. After a period of time the rate of decay of the daughter isotope becomes equal to that of the parent. This situation is known as secular equilibrium. In a radioactive decay series consisting of a very long-lived parent and a series of short-lived intermediate daughters, the condition of secular equilibrium is propagated through the entire series. The ultimate result of equilibrium is that the production of the stable daughter (an isotope of Pb) is exactly equal to the decay rate of the radioactive parent (U or Th).

U and Th are members of the actinide group of elements, which includes elements having atomic numbers from 89 to 103. These elements are chemically similar so U and Th can easily substitute for each other in minerals. U and Th are principal elements in some minerals that occur primarily in ore deposits. Many rock-forming minerals contain U or Th as trace elements, but their concentration usually is only a few parts per million. Some minerals contain larger amounts of U and Th; zircon is the most common of these minerals. Zircon also contains very low amounts of initial Pb so the U–Pb and Th–Pb methods can be applied to zircon and some other minerals using Eq. (6). One can calculate three independent ages from the three U–Pb and Th–Pb decays. If the ages agree, they are called concordant, and this is the age of the mineral. Commonly, however, the three ages are discordant or do not agree. Pb becomes volatile on heating and can be lost if the mineral is reheated at some later time. Although zircon partitions against inclusion of initial Pb there usually is a small amount of initial Pb present and larger amounts in other minerals. Thus, Eq. (7) is the appropriate age equation for the U–Pb and Th–Pb methods. Finally, the emission of an alpha particle associated with radioactive decay of U, Th, and intermediate daughters is accompanied by significant amounts of energy that can cause damage to the crystal structure of minerals. This integrated damage over time can allow Pb to escape from a mineral. The problems of initial Pb and continuing Pb loss make straightforward direct age calculations difficult or impossible. The application of isochron diagrams and a special diagram for U–Pb data, called the concordia diagram, provide powerful mechanisms to confront these problems.

The isochron diagram was described above in the section about the Rb–Sr method. The U–Pb concordia method differs from simple isochron methods in that it utilizes the simultaneous decay and accumulation of two parent–daughter pairs: ^{238}U–^{206}Pb and ^{235}U–^{207}Pb. If we rewrite Eq. (5) in terms of ^{238}U and its final daughter product ^{206}Pb:

$$^{206}\mathrm{Pb} = {}^{238}\mathrm{U}(e^{\lambda t} + 1) \tag{14}$$

and rearrange it in terms of ratio of daughter to parent:

$$^{206}\mathrm{Pb}/{}^{238}\mathrm{U} = e^{\lambda t} - 1, \tag{15}$$

where λ is the decay constant of ^{238}U (Table I).

The comparable equation for ^{235}U and ^{207}Pb is

$$^{207}\mathrm{Pb}/{}^{235}\mathrm{U} = e^{\lambda t} - 1, \tag{16}$$

where λ is the decay constant of ^{235}U (Table 1).

Various values of t may be substituted into Eqs. (15) and (16), and the resulting ratios ^{206}Pb/^{238}U and ^{207}Pb/^{235}U may be graphed (Fig. 4). These ratios for all values of t plot on a single curve called *concordia*, which is the locus of all concordant U–Pb ages. A U-bearing mineral, at the time of crystallization, contains no radiogenic Pb and the system plots at the origin. Subsequently, as long as the system remains closed to gain or loss of U and all of its daughters the system will move along the concordia curve. The age

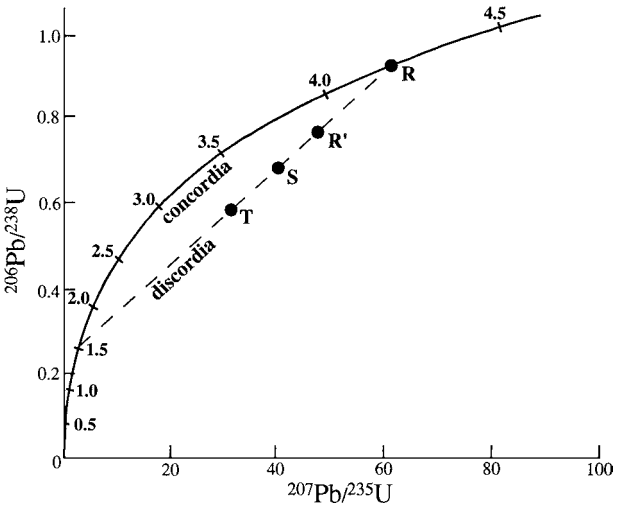


FIGURE 4 U–Pb concordia diagram. Concordia is locus of all points representing equal $^{207}Pb/^{235}U$ and $^{206}Pb/^{238}U$ ages. Location of a sample on concordia is a function of age; point R is at 4.2 Ga. Episodic loss of Pb moves a point, R, off concordia along a straight line called discordia connecting R and R′. S and T are other samples from same geologic unit or mineral, such as zircon, from same rock that have different U/P ratios. Lower intersection of discordia and concordia is time of Pb loss.

of the system at any time is indicated by its position on the curve. As long as the system remains closed, it lies on concordia and its U–Pb ages are concordant.

Let us consider a system that suffers an episode of Pb loss (or U gain) as a result of metamorphism or chemical weathering. Loss of radiogenic Pb causes the coordinates of the system to change going along a line connecting a time on concordia and the origin, that is toward both $^{206}Pb/^{238}U = 0$ and $^{207}Pb/^{235}U = 0$. This line is called a discordia line. In most geological situations, however, discordia lines do not project to the origin, but instead project to an intersection with concordia at a younger age (Fig. 4). This procedure, often called the concordia–discordia method, was devised by G. W. Wetherill in 1956. He concluded that the lower intersection of discordia with concordia was the time at which episodic Pb loss occurred from a system. The time of episodic Pb loss would be a younger reheating or metamorphism. This simple, but powerful method is used widely to interpret discordant U–Pb systems, particularly for data from the mineral zircon. Two other models used to interpret discordant U–Pb zircon ages are the continuous diffusion Pb loss model of G. R. Tilton and Pb loss related to radiation damage of zircon crystals proposed by G. J. Wasserburg. These two models use a more complex discordia curve.

Zircon is the most widely used mineral for U–Pb dating. Analyses are made using isotope dilution and a thermal-emission mass spectrometer. Samples are loaded in solid form and heated in the source region of a mass spectrometer to ionize U and Pb atoms. After correcting for the presence of nonradiogenic, or common Pb, U/Pb ages may be calculated using Eqs. (15) or (16).

In the past 20 years secondary ion mass spectrometry (SIMS) using an ion microprobe has revolutionized U/Pb dating of zircon. In an ion microprobe the sample is bombarded with a focused ion beam (typically oxygen or cesium) a few μm in diameter to produce a plasma of the target material. The secondary ionic species in this plasma are introduced into an energy filter and a mass spectrometer and measured by a sensitive counting device. The plasma contains a large variety of molecular secondary ions along with atomic ions, and a large mass spectrometer with high resolution is required for isotopic analyses of heavy elements like Pb. Analyses of 30 micron spots on zircons have permitted much more reliable interpretations of geologic history because cores of zircon crystals, secondary growth rims, and zoning can be dated independently.

F. The Radiocarbon Method

Carbon-14 is a short-lived radioactive nuclide compared to the longer-lived radioactive isotopes described previously (Table I). But, ^{14}C (radioactive carbon or radiocarbon) dating is so important to archaeology, anthropology, geology, and other fields that the method merits discussion. ^{14}C is continually produced in the upper atmosphere by interactions of cosmic ray neutrons with ^{14}N. The cosmic ray neutrons interact with the stable isotopes of nitrogen, oxygen, and carbon, but the interaction with stable ^{14}N is the most important of these reactions. The reaction of a neutron with ^{14}N produces ^{14}C and a proton is emitted from the nucleus. The atoms of ^{14}C are oxidized within a few hours to ^{14}CO, which has an atmospheric lifetime of several months. The ^{14}CO is in turn oxidized to $^{14}CO_2$. These molecules of CO_2 have a relatively long atmospheric lifetime of $\sim$100 years which allows the $^{14}CO_2$ to be well mixed and achieve a steady-state equilibrium in the atmosphere. This equilibrium is maintained by production of ^{14}C in the atmosphere and continuous radioactive decay of ^{14}C. Molecules of $^{14}CO_2$ enter plant tissue as a result of photosynthesis or by absorption through the roots. The rapid cycling of carbon between the atmosphere and biosphere allows plants to maintain a ^{14}C activity approximately equal to the activity of the atmosphere. However, the isotopes of carbon are fractionated by physical and chemical reactions that occur in nature. This fractionation introduces small systematic errors in radiocarbon dates. In addition to ^{14}C, carbon includes two stable isotopes, ^{12}C and ^{13}C. The mass-dependent fractionation can be eliminated by measuring the $^{12}C/^{14}C$ ratio on a mass spectrometer. Animals that feed on the plants also acquire

a constant level of radioactivity due to ^{14}C. When the plant or animal dies, the absorption of ^{14}C from the atmosphere stops and the activity of ^{14}C decreases due to radioactive decay. ^{14}C undergoes β-decay to ^{14}N with a half-life of 5730 years.

At some time after the death of an organism the activity of ^{14}C in dead tissue can be compared with the activity of ^{14}C in presently living tissue to yield a carbon-14 or radiocarbon date for the sample. Note that the radiocarbon method is completely different from the accumulation clocks described above. Those clocks are based on the accumulation of radiogenic daughter isotope produced by radioactive decay of a parent isotope. The radiocarbon method is based on the amount of ^{14}C remaining after radioactive decay of ^{14}C. The radioactivity of carbon extracted from plant or animal tissue that died t years ago is given by:

$$A = A_0 e^{-\lambda t}, \quad (17)$$

where A is the measured activity of ^{14}C in the sample in units of disintegrations per minute per gram of carbon, and A_0 is the activity of ^{14}C in the same sample at the time the plant or animal was alive. The carbon-14 age of a sample containing carbon that is no longer in equilibrium with ^{14}C in the atmosphere or hydrosphere is obtained by solving equation 17 for t:

$$\log_e(A/A_0) = -\lambda t \quad \text{or} \quad t = 1/\lambda \log_e(A_0/A). \quad (18)$$

The carbon-14 dating method depends on special assumptions regarding A_0 and A. These are (1) that the rate of ^{14}C production in the upper atmosphere is constant and has been independent of time, and (2) that the rate of assimilation of ^{14}C into living organisms is rapid relative to the rate of decay. As will be shown below the first assumption is not easily satisfied. It is known that the neutron flux increases with altitude above the earth's surface and that the flux is about four times greater in polar areas than at the equator. However, the ^{14}C activity is known to be independent of latitude.

The first ^{14}C ages were measured in the 1940s on elemental carbon in amorphous form ("carbon black"). However, the production of massive amounts of artificial radiocarbon from atmospheric testing of nuclear weapons in the 1950s complicated the use of elemental carbon for low-level ^{14}C measurements. Methods were subsequently developed to count the decay of ^{14}C chemically converted to purified CO_2, hydrocarbon gases and liquids. Modern laboratories can measure ^{14}C ages on organic matter as old as 40,000 to 50,000 years. A few laboratories have developed the capability to measure ages as old as 70,000 years on larger samples. In the 1970s the advent of accelerator mass spectrometry resulted in a major boost in detection efficiency. The amount of carbon required for a measurement was reduced from grams to milligrams, and the counting time was reduced from days or weeks to minutes. It was thought that the increased detection sensitivity might extend the maximum age datable by the radiocarbon method from the routine 40,000 to 50,000 years to perhaps 100,000 years. However, it turned out that contamination by younger or modern carbon, which commonly is introduced by chemical or biological activity subsequent to the death of the sample and also during sample preparation limits accelerator mass spectrometry ages to the same 40,000 to 50,000 years. It has not been possible to date with high precision samples less than about 300 years old except under special conditions. The natural fluctuations in ^{14}C production combined with the release of large quantities of fossil fuel CO_2 and production of "bomb" ^{14}C from atmospheric nuclear testing have made the measurement of young radiocarbon ages very difficult.

In the 1950s discrepancies between radiocarbon ages and true ages were noted. A long-term trend with superimposed shorter-term deviations in the ^{14}C time scale indicated that the assumption of constant production rate of ^{14}C in the atmosphere probably was not true in detail. The amount of offset between ^{14}C ages and calendar ages has been calibrated for the past 11,800 years by measuring ^{14}C ages on wood for trees for which true ages could be determined by counting yearly growth rings. By convention ages are referred to AD 1950 which is equivalent to 0 years BP (before present). From 11,800 to 24,000 years BP radiocarbon ages have been calibrated against uranium–thorium disequilibrium ages of corals or varve-counted marine sediments. The discrepancy between uncorrected ^{14}C years and calendar years at 24,000 years is 3,700 years. Computer programs are available to calculate the offset between ^{14}C and calendar years.

G. Accuracy of Radiometric Dating

Some of the geological factors that can partly or totally reset radiometric ages were outlined above. There are other factors that affect the accuracy of radiometric ages, and these must be kept in mind when evaluating radiometric ages.

The decay constants listed in Table I have been determined by direct counting experiments in the laboratory. Most are known to within an accuracy of 2% or better. The uncertainty of a decay constant generally is not included in the estimated error of the age of a rock or mineral because this uncertainty is the same for any ages measured with a given dating method. The isotopic compositions of elements used as constants in the age equations, for example, the $^{235}U/^{238}U$ ratio and the isotopic composition of atmospheric Ar, have been measured to accuracies of better than 1% and do not contribute significant errors to

age calculations. The currently accepted decay constants are used in nearly every laboratory worldwide. The last review and evaluation of decay constants and isotopic compositions for the principal dating methods was made in 1977. As new counting experiments are made or isotopic compositions are remeasured the status of individual decay constants can be evaluated. However, in general the changes are small and new decay constants are not adopted immediately because the confusion resulting from continual adoption of new constants is not worth the trouble. In 1977 values were recommended for the decay constants of ^{40}K, ^{87}Rb, ^{232}Th, ^{235}U, and ^{238}U and for the isotopic compositions of atmospheric Ar and K, Rb, Sr, and U.

Modern analytical instruments, particularly mass spectrometers used for isotopic measurements, permit measurement of isotopic ratios to a few tenths of a percent or better. This does not mean the accuracy of an age measurement is as good as this because other factors, such as the calibration of isotope dilution tracers, affect the accuracy of an age determination.

Finally, it is important to realize that geochronologists do not rely entirely on error estimates or fits to lines on correlation diagrams to evaluate the accuracy of radiometric ages. The simplest way to check the reliability of an age measurement is to repeat the analytical measurements in order to minimize analytical errors. This procedure helps to minimize human error and also provides information on which to determine analytical precision. Another technique is to measure ages on several samples (rocks or minerals) from the same rock unit. This technique helps to identify geological factors because different minerals respond differently to heating and chemical changes. Stratigraphy also provides an important constraint where overlying and underlying rocks units have been dated. Finally, the use of different decay schemes on the same rock is a very good way to check the accuracy of measured ages. If two or more radiometric clocks, that is, different radiometric dating methods, running at different rates give the same age, this is powerful evidence that the ages are correct.

SEE ALSO THE FOLLOWING ARTICLES

GEOLOGIC TIME • RADIOCARBON DATING • RADIOMETRY AND PHOTOMETRY • THERMOLUMINESCENCE DATING

BIBLIOGRAPHY

Dalrymple, G. B., and Lanphere, M. A. (1969). "Potassium-Argon Dating," W. H. Freeman and Company, San Francisco.

DePaolo, D. J. (1988). "Neodymium Isotope Geochemistry," Springer-Verlag, Berlin.

Doe, B. R. (1970). "Lead Isotopes," Springer-Verlag, New York.

Faure, G. (1986). "Principles of Isotope Geology," 2nd ed., Wiley, New York.

Faure, G., and Powell, J. L. (1972). "Strontium Isotope Geology," Springer-Verlag, Berlin.

Jäger, E., and Hunziker, J. C. (eds.) (1979). "Lectures in Isotope Geology," Springer-Verlag, Berlin.

McDougall, I., and Harrison, T. M. (1999)."Geochronology and Thermochronology by the $^{40}Ar/^{39}Ar$ Method," 2nd Ed., Oxford Univ. Press, New York.

Radionuclide Imaging Techniques, Clinical

Peter F. Sharp
University of Aberdeen

GLOSSARY

Compton scatter Interaction between a γ ray and an atom in which the γ ray is scattered and also loses some of its energy by ejecting an electron from the atom.
Electron capture Radioactive decay process in which one of the orbital electrons is captured by the nucleus. A γ ray is emitted from the nucleus but without the need for a charged particle to be emitted as well.
Isomeric transition Radioactive decay process in which a γ ray is emitted without an accompanying β particle.
Photomultiplier tube Device that converts light into an electronic signal. The strength of the signal is proportional to the intensity of the light.
Positron Particle having the same mass as an electron but with a positive rather than a negative charge. When emitted from the nucleus of a radioactive element it interacts with an electron, the two particles disintegrating to form two γ rays.
Radionuclide Radioactive element.
Radiosensitive Tissues that are particularly prone to damage if hit by radiation.
Scintillation crystal Transparent crystal with the property of emitting light when hit by a γ ray.

CHEMICAL RADIONUCLIDE IMAGING obtains diagnostic information by observing the *in vivo* distribution of a pharmaceutical administered to the patient. This is achieved by attaching a radioactive label to the pharmaceutical and producing an image from the radiation emitted from the body. Due to the difficulties in imaging γ radiation, image quality is significantly worse than in X radiography. This is offset, however, by the ability of the technique to permit the observation of specific physiological processes depending upon the pharmaceutical employed.

I. INTRODUCTION

The potential for radiation to damage human tissue is well known, and so it is perhaps surprising to find that it is widely used for the diagnosis of a variety of diseases ranging from cancer to dementia.

The earliest medical potential for radioactivity was seen to lie in its longevity of action, the naturally occurring radioisotopes known at that time had half-lives of thousands of years. These materials were used for treating a wide variety of conditions, such as gout and rheumatism, but the eventual fate of the patients in these early trials is not recorded.

When the potentially destructive effect of radiation on tissue was appreciated, the value of radiation for destroying tissue without the need for surgery was obvious and the technique of radiotherapy became widely used.

If low levels of radioactivity are introduced into the body it becomes possible to detect the emerging radiation and so determine the distribution of the radioactivity *in vivo*. However, to provide diagnostic information it is first necessary to choose a material whose distribution in the body may provide information of clinical relevance. The radioactive label is then attached to this pharmaceutical to act as a purely passive tracer, and the distribution of the pharmaceutical provides the diagnostic information.

The diagnostic value of radionuclide imaging thus depends on the selection of an appropriate radioactive label, the choice of an effective pharmaceutical, and the ability to produce an image from the emerging flux of radiation.

II. RADIOPHARMACEUTICAL

A. Choice of Radioactive Label

The properties required for a radioactive label to be effective are

1. its half-life should be similar to the length of the test,
2. the radionuclide should emit γ rays and preferably no charged particle should be emitted,
3. the energy of the γ rays should be between 50 and 300 keV,
4. the radionuclide should be readily available at the hospital site, and
5. the radionuclide should be chemically suitable for incorporating into a pharmaceutical without altering its biological behavior.

The main requirement of the label is that it emit γ rays since this is the only type of radiation that can be readily detected externally while not causing significant damage to the patient's tissue. The energy of the γ ray must also be sufficiently high for it to penetrate through the tissue; in practice this means that the minimum energy should be about 50 keV. If the energy is too high it will become difficult for the γ ray to be stopped in the detector system. This puts an upper limit of about 300 keV, with a value of about 100 keV being ideal.

Ideally the radionuclide should decay by the emission of a γ ray with no α or β particles as these have a very short range in tissue and contribute to tissue damage. In practice, the ratio of γ rays to charged particles can be maximized by using radionuclides that decay by isomeric transition or electron capture. The radiation emitted when a positively charged β particle, a positron, disintegrates has also been used for imaging; this is the so-called positron emission tomography (PET).

Obviously the likelihood of damage increases the longer the radioactive material remains in the patient. A radionuclide with a half-life comparable to the length of the study is ideal (i.e., a few hours).

This latter requirement conflicts with the need to ensure a regular supply of material to the hospital. Radionuclides can be produced from either a nuclear reactor or a cyclotron. Neither source is particularly convenient for medical usage, although some hospitals in the United States have medical cyclotrons. A third and the principal source of radionuclides is from generators. These consist of a radionuclide with a long half-life that decays into the short-lived material needed to label the pharmaceutical. The short-lived daughter product is chemically separated from the long-lived parent when required. As the parent continues to decay a new supply of the daughter produce is generated. If the daughter is not eluted from the parent, then it apparently decays with a half-life equal to that of the parent. The generator thus succeeds in the apparently impossible task of providing a short-lived radionuclide from a source with a long half-life. The most commonly employed generator has a molybdenum parent producing a daughter product of technetium 99*m*. Technetium 99*m* (^{99m}Tc) is the radiolabel used in most clinical studies. It has a half-life of 6 h and emits a γ ray of 140 keV energy by isomeric transition.

B. Choice of Pharmaceutical

One factor has not yet been considered, namely, the need for the radionuclide to be chemically suitable for incorporating into a pharmaceutical without changing its biological kinetics. This is not a trivial problem as the elements most suitable chemically, such as carbon and nitrogen, do not emit suitable radiation. It is due to the ingenuity of the radiochemist that it has been proven possible to label

TABLE I Commonly Used Radiopharmaceuticals

Radioactive label	Pharmaceutical	Organ to be studied	Clinical application
^{99m}Tc	Pertechnetate	Brain	Metastases, infarction
^{99m}Tc	Macroaggregates	Lung	Embolus
^{99m}Tc	Colloid	Liver	Metastases cirrhosis
^{99m}Tc	Methylene diphosphonate (MDP)	Bone	Metastases
^{99m}Tc	Diethylene triamine pentaacetic acid (DTPA)	Kidneys	Renal function and drainage
^{99m}Tc	Red blood cells	Heart	Cardiac wall motion and cardiac volume
^{99m}Tc	White blood cells	Whole body	Sites of infection and inflammation
^{99m}Tc	Hexamethyl propyleneamine oxime (Ceretec)	Brain (cerebral blood flow)	Dementia, infarction
^{99m}Tc	Imminodiacetic acid (IDA)	Hepatobilinary system	Biliary obstruction
^{201}TI	Thallous chloride	Heart	Infarction, ischemia
^{123}I	Iodide	Thyroid	Function
^{123}I	Hippuran	Kidneys	Renal function and drainage

a wide range of pharmaceuticals with elements that appear at first sight to be most unsuitable, such as ^{99m}Tc. The commonly used radiopharmaceuticals are listed in Table I together with the diagnostic test for which they are used. Apart from the ability to label it with a suitable radionuclide, the effectiveness of the pharmaceutical depends also upon its specificity for the particular biological feature it is required to study. The distribution of the radiopharmaceutical to organs other than those targeted often limits the amount of radioactivity that can be administered, particularly if these organs are radiosensitive. Finally, to minimize the dose of radiation received it is desirable that the residence time of the radiopharmaceutical in the body should be similar to the time required to carry out the study.

III. GAMMA CAMERA

The principle of the gamma camera was proposed by Hal Anger in 1958, and while it has become a more sophisticated device, the basic concept has remained unchanged. It is now used almost exclusively for imaging in nuclear medicine.

Figure 1a shows a typical camera. The detector head cross section shown in Fig. 1b consists of a large scintillation crystal, between 40 and 50 cm in diameter, made from sodium iodide to which trace quantities of thallium have been added [NaI(Tl)]. Between the crystal and the patient is a collimator. In the most commonly used form it consists of a lead plate with several thousand parallel holes running through it. This collimator performs

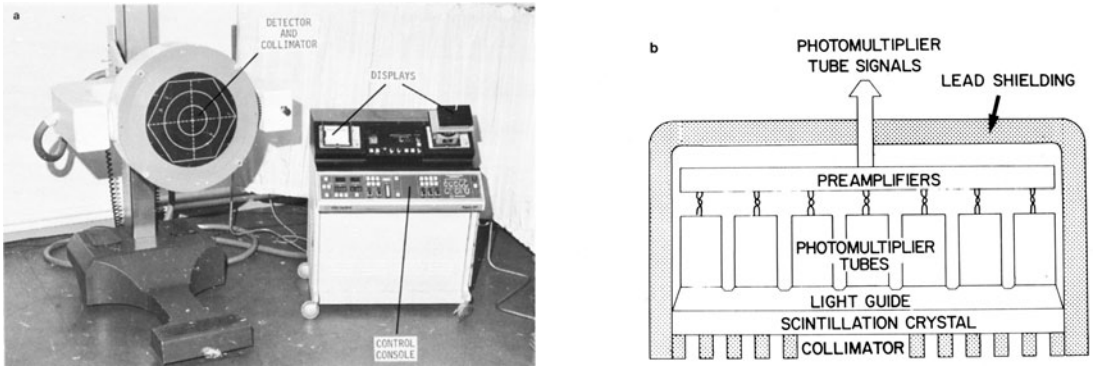


FIGURE 1 The gamma camera. (a) A typical gamma camera. (b) Cross section through the detector head. [Reproduced with permission from P. F. Sharp, P. P. Dendy, and W. I. Keyes (1985). "Radionuclide Imaging Techniques." Academic Press, London, UK.]

a function similar to that of a lens in an optical camera. Since the energy of γ radiation is considerably higher than that of light it is not possible to image the γ rays by focusing them on the crystal, instead the image is produced by excluding all γ rays except those traveling along the axis of a hole. Rays which are obliquely incident on the collimator are absorbed as they try to pass through the lead septa between the holes. The collimator, thus, images by cutting out all γ rays except those traveling in a specific direction. One obvious consequence is that only a very small proportion of the emitted radiation will be used to form the image; the camera will have a very low sensitivity. Typically only 0.1% of the γ rays emitted during the imaging time will form the final image. A second problem is that both sensitivity and image sharpness will depend on the dimensions of the collimator holes. Increasing the diameter of each hole will increase sensitivity but only by allowing through some γ rays obliquely incident on the collimator. This results in the sharpness or spatial resolution of the camera deteriorating.

Resolution also worsens as the distance between the collimator and the patient increases, so it is necessary to get the patient as close as possible to the collimator. Under near-optimum imaging conditions a spatial resolution of between 7 mm and 1 cm can be achieved. Formally this means that two small points of radioactivity this distance apart could just be distinguished.

The imaging ability of the collimator depends upon the septa being thick enough to absorb any γ rays hitting them. Thus, collimators are designed for use with γ rays of either low ($<$150 keV), medium ($<$250 keV), or high ($<$400 keV) energy. A series of collimators will usually be available with each camera, the most appropriate one depending upon the energy of the γ ray and whether high sensitivity, good spatial resolution, or a compromise between the two is required for the study.

The image of the radiopharmaceutical distribution produced in the crystal is made up of brief flashes of light, the scintillations resulting from the γ rays being converted into visible light by the crystal. Each scintillation is then converted into electronic signals by an array of photomultiplier tubes (PMT) situated at the rear of the scintillation crystal. Typically there are between 37 and 61 tubes arranged in a close-packed hexagonal array to ensure maximum coverage of the crystal. The output signals from the PMTs are then processed to give the spatial position of each scintillation. This is done by modifying the output from each PMT by a factor which is unique to the tube and then summing it with modified signals from the other tubes. Each PMT has two factors associated with it, one set being used to produce the x coordinate the other the y coordinate of the scintillation.

There is also a variation in the energy signal depending upon where the γ ray interacts in the crystal. Both spatial nonlinearity and spatial variation in the energy signal are corrected, to a significant extent, by microprocessor controlled circuitry. However, image blurring cannot be corrected.

Although the collimator excludes γ rays according to the direction in which they are traveling when they arrive at the collimator, this does not preclude the possibility that they might have been scattered before arriving at the collimator. In practice the probability of a γ ray undergoing Compton scattering is high. As Compton scattering involves a loss of energy by the γ ray it is possible to discriminate between scattered and unscattered radiation by measuring the energy of the detected γ ray. This is easily done since the amount of light in the scintillation is proportional to the energy deposited in the crystal by the γ ray. The sum of the (unmodified) output signals from the PMTs constitutes the energy signal that is then sent to a pulse height analyzer. This device will only pass signals whose height (i.e., energy) falls between values selected by the operator.

The X and Y signals are applied to the deflection plates of the cathode ray tube (CRT) display so locating the tube's beam at the same position as that of the detected scintillation in the crystal. If the energy signal is found to be in the expected range, then an unblank signal is passed to the CRT brightening up the beam so that a spot of light appears briefly on the display. These flashes of light are integrated on photographic film. The final image, thus, appears as a series of small spots, each one representing one detected γ ray (Fig. 2).

The image suffers some loss of quality in the transfer from crystal to display. There is some image blurring, although compared to that introduced by the collimator it is small, and a loss of spatial linearity (i.e., the image is distorted). This loss of image quality is mainly the result of creating an image from signals that are weak and using a small number of PMTs. Increasing the number of tubes does not help since it would reduce even further the amount of light received by each one. In the modern camera, microprocessors allow correction to be made for

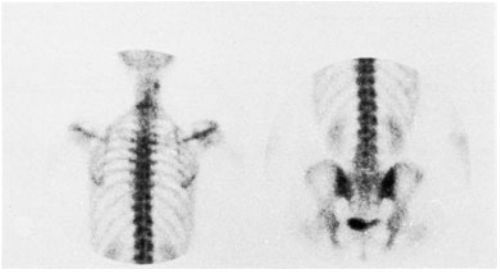


FIGURE 2 Two gamma camera images showing the distribution of the bone-seeking radiopharmaceutical ^{99m}Tc MDP. Each detected γ ray is displayed as a small dot.

the non-linearity. It is also necessary to correct the energy signal to ensure that its value does not depend on where the γ ray interacts in the crystal.

The gamma camera is also restricted as to the maximum rate at which it can handle γ rays. When the input count rate becomes too high, the camera fails to distinguish consecutive γ rays as separate events. The result is a loss of linearity between detected and recorded count rate and a distortion of the image. The maximum count rate is about 30,000 counts per second, which is good enough for all but a few specialized studies.

IV. TOMOGRAPHY

The gamma camera produces two-dimensional images of a three-dimensional distribution of the radiopharmaceutical. Consequently information about the depth of an organ is missing, and this is usually obtained by taking additional views from different angles. True three-dimensional images can, however, be produced by a tomographic imaging technique referred to as single photon emission computed tomography (SPECT). This consists of taking a series of views, typically 64, as the gamma camera is rotated around the patient (Fig. 3a). This data is then processed by a computer to produce images representing slices of the radiopharmaceutical distribution usually in one of three orthogonal planes; the plane of rotation of the camera being known as the transaxial, the other two are the coronal and sagittal.

Each row of a single camera image can be thought of as consisting of a single view of the transaxial plane; this view represents, in fact, the count density pattern in this plane projected onto the row. The same row on each of the 64 images provides information about the same transaxial plane but projected from different directions. To reconstruct the three-dimensional image it is necessary to backproject this data, that is, it is assumed that all points in the plane have the same value as at the corresponding point on the row. Backprojecting the data from the angle at which the image was acquired results in a transaxial slice being produced (Fig. 3b). While this simple reconstruction algorithm gives an approximate image of the plane it does suffer from several distortions; in particular, a high background count density and a blurring of the edges of structures. To overcome this the data is filtered or processed either before or after the backprojection.

The complete set of transaxial sections, one from each row of the planar views, provides a cube of data from which sections can be extracted in the coronal and sagittal directions and in any oblique plane (Fig. 3c).

The advantage of SPECT is not only that it gives full three-dimensional information about the position of image

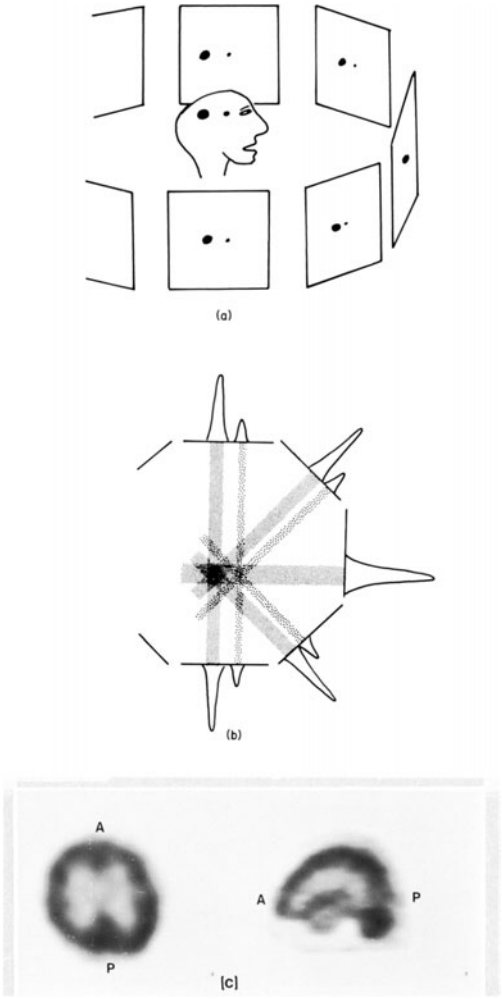


FIGURE 3 Single photon emission tomography. (a) The gamma camera is rotated around the patient and a series of images are taken from different angles. (b) To reconstruct a transaxial slice a row of data in each image is backprojected. The data in each row is shown in profile; the height of the curve represents the count density in the image. Superimposition of this backprojected data produces the image, in this case the two areas where the radiopharmaceutical has concentrated. (c) A transaxial and a sagittal section through the brain. This radiopharmaceutical demonstrates cerebral blood flow. A indicates anterior and P posterior.

structures but that it also improves image contrast, the masking effect of radioactivity in planes above and below that of interest having been removed. The expectation that it would allow the precise quantitation of the amount of radiopharmaceutical present in an organ has not been fulfilled. The main problem is that radiation is attenuated before reaching the camera to an extent that depends on the nature of the structures through which it has to pass. Unfortunately it is only the tomographic section that can provide this information, yet it is needed before the section can be

reconstructed! Several attenuation correction algorithms have been proposed but none provide any great accuracy.

While the technique of SPECT has limitations, it does overcome several of the degradations inherent in planar imaging, yet for many years it was a technique that was looking for a clinical application. This problem was partly overcome when it became clear that gamma cameras capable of carrying out both tomography and conventional imaging could be purchased at very little extra cost. Also in the last few years radiopharmaceuticals have been produced that require tomography for effective interpretation.

Systems designed specifically for tomography are available, and in particular, they overcome some of the problems with lack of sensitivity found in the rotating camera. They may be a conventional camera system but with two or three detector heads or consist of a ring of smaller detectors. This latter system will produce only a limited number of tomographic slices. These systems are, however, expensive, and the ring detector systems are only suitable for tomography. It is unlikely that they will be used other than in specialized centers.

The technique discussed so far has used single γ-ray photons. Reference was made earlier to the use of the γ radiation given off when a positron annihilates. This takes the form of two γ rays that travel in almost exactly opposite directions. If the simultaneous arrival of these two γ rays in a detector that surrounds the patient is noted, then the point of emission of the positron can be assumed to lie somewhere along the line between the two interactions. By looking for points of intersection of many of these pairs of lines, a three-dimensional image can be produced without the need to employ collimators (Fig. 4a). The technique requires specialized imaging equipment (Fig. 4b) using rings of detectors around the patient. Positron-emitting radionuclides can be produced from a generator, but the most interesting ones require an on-site cyclotron. Positron-emitting radionuclides do, however, include some of the most biologically interesting elements, such as carbon, oxygen, and nitrogen, and a whole range of interesting studies are possible. In comparison with SPECT it is very expensive and time consuming to perform and, at least in the near future, will only be used in the most advanced centers.

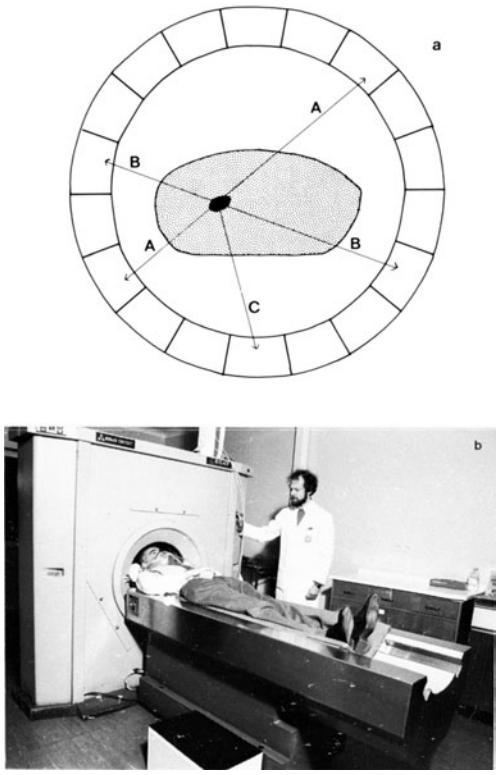


FIGURE 4 Positron emission tomography. (a) The decay of a positron produces pairs of γ rays that travel in opposite directions, such as rays *AA* and *BB*. They can be distinguished from single γ rays, such as *C* which is not produced from a positron, by their simultaneous arrival in the detector. The point of emission of a positron must lie along the line joining the points at which the pair of γ rays hit the detectors. Where such lines intersect gives the location of the radiopharmaceutical in three dimensions. (b) A positron imager. The patient is breathing a radioactive gas.

V. DATA PROCESSING

The imaging capability of the gamma camera can be extended by interfacing it to a computer system and acquiring the image in a digitized form. This allows the image data to be manipulated in a variety of ways (e.g., for the reconstruction of tomographic images as described in the previous section). Only two aspects of data processing will be considered here in detail, the quantitation of the data in an image and the enhancement of image presentation.

A. Quantitation

The digitized image is represented in the computer as a two-dimensional array of numbers, each of which denotes the number of γ rays acquired in that area of the image. It is, therefore, possible to make measurements of the number of γ rays acquired from a specific part of the image (e.g., from a particular organ). Unfortunately this does not allow us to deduce the precise amount of radiopharmaceutical present in the organ since some of the γ rays emitted will be absorbed or scattered in their passage through overlying tissues and so will not be recorded. As the thickness of this tissue and its attenuating ability is not known, it

is only possible to make an approximate estimate of the radiopharmaceutical concentration.

Even this limited ability to quantify the image data can, however, produce useful clinical information. It is particularly valuable in assessing how the amount of radiopharmaceutical in an organ varies during a study; from which information on how well an organ is functioning can be inferred. For example, Figs. 5a–5c show a series of images of the kidneys produced using the radiopharmaceutical ^{99m}Tc DTPA. This material is removed from the bloodstream by the kidneys, the rate at which it concentrates in the kidneys indicating renal function. The radiopharmaceutical passes through the kidneys and finally drains from them with the urine. This part of the study gives information on the drainage capability of the kidneys. A more precise assessment can be made by collecting the study as a sequence of images on the computer. Typically the study will consist of 60 consecutive images each 20 s in duration. Figure 5e shows curves generated from this data, representing the count density recorded over each kidney and the bladder as a function of time. In the abnormal study shown in Fig. 5f, the radiopharmaceutical concentrates in the kidneys but one of them drains poorly.

B. Image Presentation

The amount of information that a study provides is obviously dependent on how well the data is presented to the clinician. The quality of the acquired data may be extremely high but if the image is overexposed, for example, its interpretation will be impaired. The conventional gamma camera analog image recorded on photographic film suffers from such problems. In contrast, the digitized image can be viewed on the computer TV display and its intensity and contrast altered by the observer until he is certain that all useful information has been extracted. Computer technology also allows the image to be displayed in ways other than the conventional black and white. For example, using different colors to represent image intensity may facilitate interpretation.

The basic image data can also be modified by filtering it prior to display; smoothing filters are used to reduce image noise and sharpening filters to accentuate the edges of structures. The full potential of image filtering has not yet been realized.

The interpretation of dynamic studies can be facilitated by displaying the images in rapid succession on the screen. This gives the impression of continuous movement of the radiopharmaceutical, the so-called cine mode display. Three-dimensional displays of data are also possible although they are not yet routinely employed.

VI. FUTURE DEVELOPMENTS

A. Radiopharmaceuticals

The development of a new radiopharmaceutical can alter significantly the clinical role of nuclear medicine. But, as with any pharmaceutical, the production of new agents is a slow and expensive process. In nuclear medicine there is the further problem that it may not prove possible to predict the clinical role of a new material. For example, a radiopharmaceutical ^{99m}Tc HMPAO was developed with the aim of giving images of regional cerebral blood flow. While it had obvious applications in the detection of abnormal cerebral pathology, such as cerebral infarction, its adoption for this role depended on how effectively it competed with other established techniques. In fact, early experience showed that one of its most useful applications may be in the differential diagnosis of dementia, an area where clinical imaging techniques have had little impact in the past. Yet such as application could not have been predicted when the decision was taken to develop the material.

This same material was also found to provide a very effective way of radiolabeling white blood cells. This gives a very sensitive technique for detecting sites of infection in the body, a major clinical problem.

One of the main goals in radiopharmacy has been to develop a radiopharmaceutical that concentrated only in the organ or structure of interest. For example, if a radiopharmaceutical concentrated only in tumors and not in other tissues, it would be possible to detect them at a very early stage of development when the chance of successful treatment was high. Such a pharmaceutical could also be used to target a radiolabel that would destroy the tumorous tissue. The principle hope for many years has been on radiolabeled monoclonal antibodies. While the desired specificity has not yet been achieved there are signs that some very useful agents may be produced.

Such specificity is not restricted to the task of detecting tumors. The concept of producing materials that would localize in specific neurotransmitters or neuroreceptors is one that is attracting a lot of interest. Such radiopharmaceuticals would provide information on a variety of clinical conditions affecting the nervous system, such as Parkinson's disease.

B. Instrumentation

The design of the gamma camera has not changed significantly over the years. Novel detectors using semiconductor materials or the cheap, large-area gas-filled detector chambers have been proposed but problems have been encountered that have, so far, hindered their development as clinical instruments. It has even been shown that

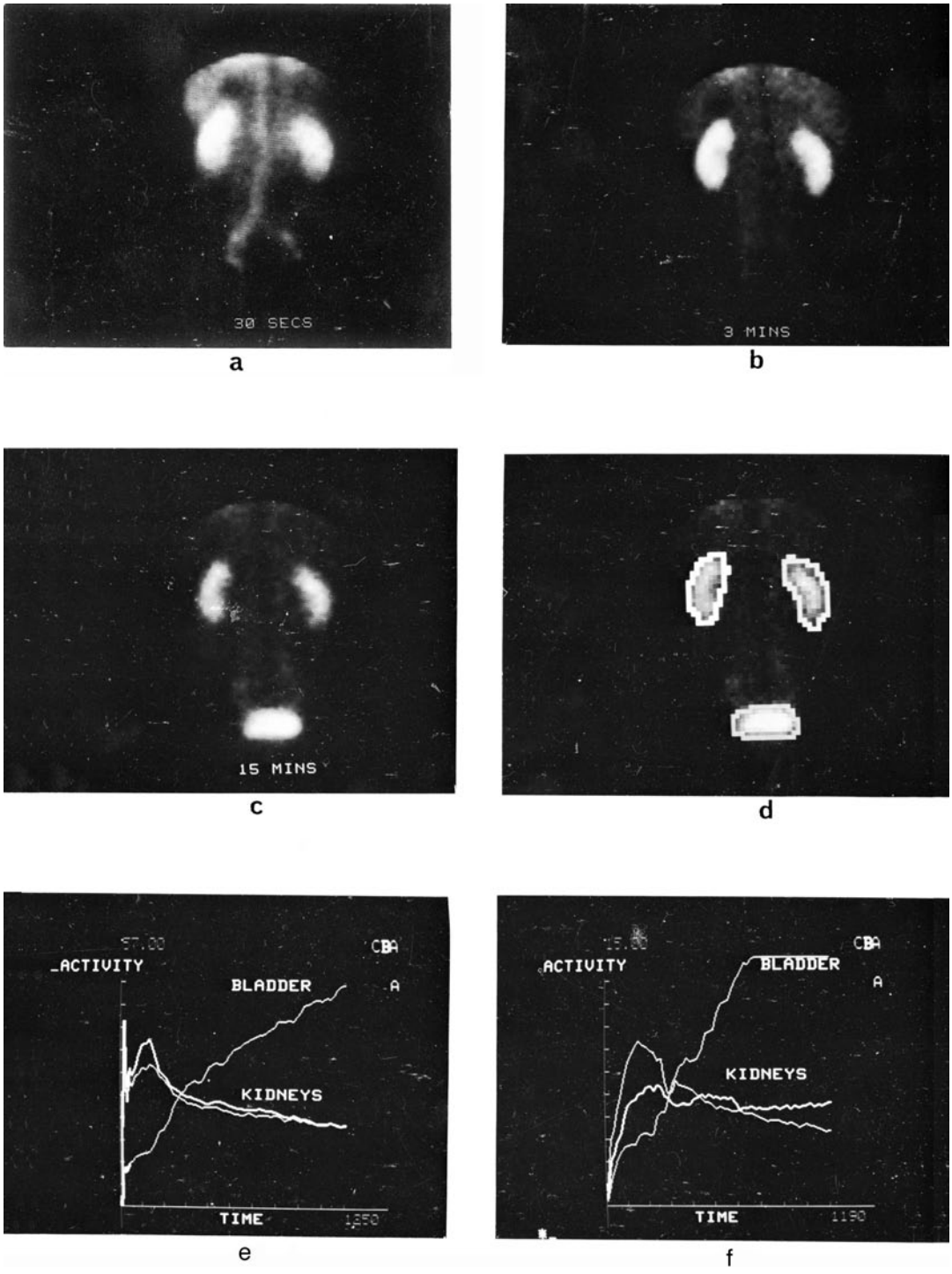


FIGURE 5 The dynamic study, (a)–(c) Images taken during the study showing the uptake of the radiopharmaceutical in the kidneys and its eventual drainage into the bladder. (d) Areas (regions of interest) have been drawn around the kidneys and the bladder. (e) Curves showing the number of γ rays inside each of the areas from each of the series of images taken during the dynamic study. These normal kidney curves show an initial rapid rise as the radiopharmaceutical concentrates in them, then a fall as it drains into the bladder. (f) In this study one of the kidneys is shown to be draining slowly.

it is possible to produce images without employing a collimator, by using coded apertures. These not only gave the promise of an imaging device with extremely high sensitivity but also one which would produce tomographic images. However, once again the idea proved to have drawbacks that have precluded its use for clinical imaging.

The gamma camera appears likely to remain the standard imaging instrument for the next few years. It has undergone some changes, in particular the use of microprocessors to correct imaging defects. The digital camera has the same detector as the conventional camera but the X, Y, and energy signals are digitized. These digital signals are then fed, after correction, directly into the camera's own computer where they are presented on a digital display. The computer is also used to control other functions of the camera such as pulse-height analysis, acquisition time, and labeling of the image with the patient's details. It is a logical step from the separate camera and computer.

SEE ALSO THE FOLLOWING ARTICLES

HEALTH PHYSICS • ISOTOPES, SEPARATION AND APPLICATIONS • PHARMACEUTICALS • RADIATION EFFECTS IN ELECTRONIC MATERIALS AND DEVICES • RADIOACTIVITY • TOMOGRAPHY

BIBLIOGRAPHY

Harbert, J., and Da Rocha, A. F. G. (eds.) (1984). "Textbook of Nuclear Medicine," 2nd ed., Vol. 1. Lea & Febiger, Philadelphia, Pennsylvania.

Jackson, P. C. (1986). "Radionuclide Imaging in Medicine," Farrand Press, London, UK.

Sharp, P. F., Dendy, P. P., and Keyes, W. I. (1985). "Radionuclide Imaging Techniques," Academic Press, London, UK.

Sorenson, J. A., and Phelps, M. E. (1987). "Physics in Nuclear Medicine," 2nd ed. Grune & Stratton, Orlando.

Scanning Electron Microscopy

David C. Joy
University of Tennessee

D. G. Howitt
University of California

GLOSSARY

Brightness Quantity of focused output from an electron source (measured in amperes per square centimeter per steradian).
Convergence Semiangle of the cone formed by the extremal rays focused on to the specimen.
Frame One complete scanned image.
Pixel Picture element. The smallest definable unit of contribution to an image.
Raster The pattern, usually rectangular, traced on the specimen by the scanning beam.
Scintillator Material which produces light under the impact of an energetic electron.
Working distance Clear space between the final defining aperture and the specimen surface.

THE SCANNING ELECTRON MICROSCOPE is a microscope that utilizes electron beams to produce high spatial resolution and high depth of field images of solid specimens. Such images provide information about the topography as well as the chemical, magnetic, and electronic state of a sample. The SEM is the most widely used electron-optical device, and has become as easy to use as an optical microscope.

I. HISTORY

The scanning electron microscope (SEM) originated in the pioneering work of Knoll and Von Ardenne in Germany in the 1930s. Von Ardenne, in particular, anticipated all of the features of the modern SEM, but he worked at a time when experimental technique was not sufficiently advanced to put these ideas into practice. A more advanced design was described by Zworykin, Hillier, and Snyder in the United States during World War II but, although highly sophisticated, the instrument produced results that were poor compared with those from the then rapidly developing transmission electron microscope, so the project was terminated. The work of Oatley and his students in Cambridge, beginning in 1948, laid the foundation of the SEM in its present form. Current commercial designs differ very little from the principles developed by the Cambridge group.

II. THE INSTRUMENT

A. Design Principles

The basic principles of the instrument are shown in Fig. 1. The SEM uses two electron beams, where both are scanned in an identical regular raster pattern. One beam is incident on the specimen to record the data and the other onto a cathode ray screen to display it. Signals of any type emitted from the specimen under the stimulation of the beam are collected by a suitable detector, amplified, and used to modulate the brightness of the display tube. The result is thus a map of the specimen, related to whatever signal is collected, rather than an image in the conventional sense. In the SEM, magnification is obtained by decreasing the areas scanned. If the raster scanned on the sample is a square of dimension $A \times A$, and the raster scanned on the display screen is $B \times B$, then a linear magnification of B/A is achieved. Since the size of the display screen is fixed, the magnification may be changed by varying the side A of the square scanned on the sample. This arrangement has several advantages:

1. The sample can be imaged from any electron-stimulated emission, even those which cannot be focused.
2. Multiple signals can be collected and displayed simultaneously.
3. Altering the magnification does not necessitate refocusing the instrument, and changing the magnification does not lead to image rotation, unlike as with other electron microscopes.
4. Since the signal is recorded sequentially as a time-varying voltage, the image can be electronically processed before it is displayed or recorded.

B. Electron–Optical Performance

The quality of the micrographs produced by the SEM is determined by a variety of factors not the least of which is the amount of signal that can be obtained from the sample.

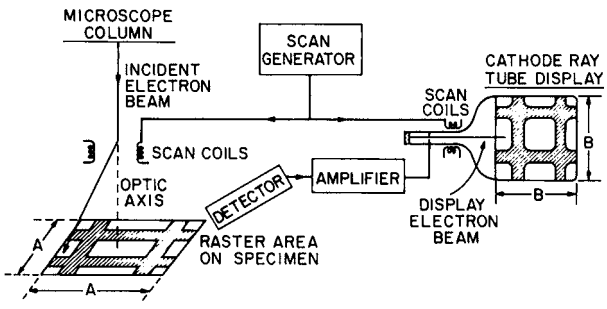


FIGURE 1 Principle of the scanning electron microscope.

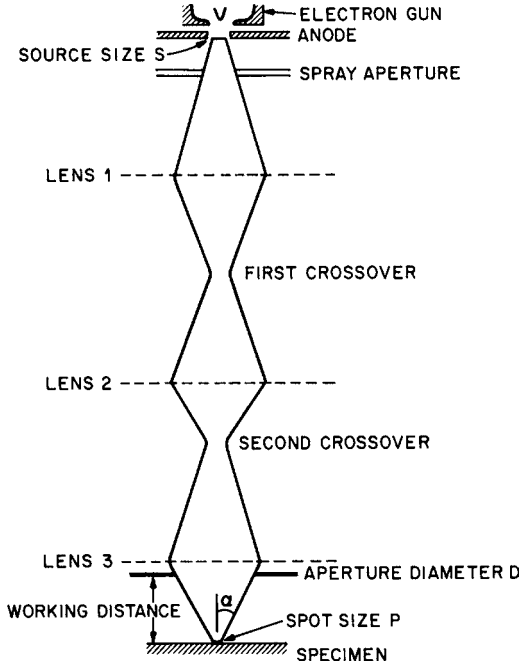


FIGURE 2 Optical ray path of the SEM.

When this is not the limiting factor the two other significant parameters are the diameter of the electron probe at the surface of the sample and the nature of the electron beam interaction which produces the emission and thereby broadens or spreads the effective size of the probe. Figure 2 shows, schematically, the optics of the SEM. Typically these consist of an electron gun, two or three electromagnetic lenses, the sample, and one or more detectors.

Typically the beam interaction takes place in a vacuum because of the ease with which electrons are absorbed in air, however, some of the more modern instruments are capable of operating at intermediate pressures down to 1/50th of an atmosphere. The advantages to the presence of a residual atmosphere include the ability to evaluate specimens *in vivo* as well as to locate reaction chambers.

The electron probe size d is given by

$$d = sM_1M_2M_3, \tag{1}$$

where s is the effective diameter of the electron source (about 50 μm for a tungsten thermionic emitter, 10 μm for a lanthanum hexaboride source, and about 10 nm for a cold field emitter), and M_1, M_2, and M_3 are the demagnifications of the condensor lenses. By varying the excitation of the lenses the beam diameter can be set to any desired value from the source size downward. The spatial resolution of the SEM is determined by the spot size at the specimen, but the resolution cannot be made arbitrarily high because of fundamental constraints on the system. First, there are aberrations in the probe-forming

optics that increase the probe diameter above the value predicted by Eq. (1). For the beam convergence angle α, defined from Fig. 2 as the aperture radius divided by the working distance, the actual probe diameter p is given as

$$p^2 = d^2 + \left(0.5C_5\alpha^3\right)^2 + (\lambda/\alpha)^2, \tag{2}$$

where C_5 is the spherical aberration coefficient of the lens, λ is the de Broglie electron wavelength equal to $12.26/(E^{1/2})$ Å, and E is the beam energy in electron volts. The value of α can be chosen to maximize the instrument performance in one of several ways. One choice would be to obtain the smallest probe size, but in practice a more important consideration is to produce the electron beam current into a probe of given size. An electron gun has a finite brightness (current density/unit solid angle) and the brightness is constant at all points in the optical system. With the restriction that the brightness B is constant it follows that a value of

$$\alpha = (p/C_s)^{1/3} \tag{3}$$

will put a maximum current, equal to

$$I = 1.88Bp^{8/3}/C_s^{2/3}, \tag{4}$$

into the probe of diameter p. For a tungsten thermionic electron source operating at 20 keV the brightness B is of the order of 10^5 A/cm^2-sr, and C_s on a modern instrument is typically 1 cm. Applying Eq. (3) shows that for desired probe sizes in the range 10 nm to 1 μm, α is in the range 1 to 10 mrad (0.05–0.5°), and the incident beam current [from Eq. (4)] is then between 10^{-12} and 10^{-6} A. More advanced electron sources, such as field emission guns, have a brightness which is 100 to 1000 times greater than that of a tungsten hairpin gun. Such guns are now becoming widely used in high performance SEMs because they permit both higher spatial resolution and larger probe currents.

The depth of field D_f of the image, defined as the vertical distance between the points at which image resolution degrades beyond the expected value, is given as

$$D_f = (\text{pixel size})/\alpha, \tag{5}$$

where the pixel size is the effective value at the specimen. Since α is between 10^{-3} and 10^{-2} (rad), the depth of field is thus several hundred times the pixel size. At low magnifications the depth of field can therefore approach millimeters in extent, giving the SEM an unrivaled ability to image complex surface topography, and produce images with a pronounced three-dimensional quality to them.

C. Detection Limits

Information in the SEM image is conveyed by changes in the magnitude of the detected signal. If the average signal level is S, and if some feature on the sample causes signal changes δS, then the feature is said to have a contrast level C given by

$$C = \delta S/S. \tag{6}$$

Changes in the signal also occur because of statistical fluctuations in the beam current, and in the efficiency with which the various electron–solid interactions occur. These statistical variations constitute a "noise" contribution to the image, which therefore has a finite signal-to-noise ratio. For image information to be visible the magnitude of the signal change occurring at the specimen must exceed the magnitude of the random variations by a factor of 5 to 10 times. This leads to the concept of a threshold current I_{th}, which is the minimum beam current required to observe a feature with contrast C.

$$I_{th} = 4 \times 10^{-12}/\left(C^2 t_f\right) \text{ A}, \tag{7}$$

where t_f is the total time taken to scan the image (assumed to contain 10^6 pixels). The observation of low contrast features therefore requires high beam currents or long collection times.

The threshold current requirement sets a fundamental limit to the performance of the SEM in all modes of operation, and it is in most cases that the image detail is predicted by lack of signal rather than by microscope performance.

III. MAJOR ELECTRON–SOLID INTERACTIONS

The interaction of the electron beam with a solid specimen produces a wide variety of emissions, all of which are potentially useful for imaging (Fig. 3). Each of these signals is produced with a different efficiency, comes from a different volume of the sample, and carries different information about the specimen. In the following section the major interaction mechanisms which produce electrons

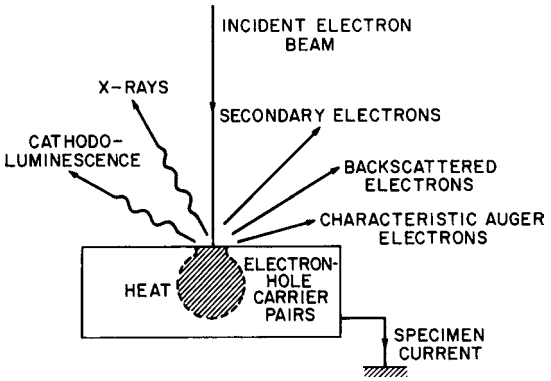


FIGURE 3 Possible electron–solid interactions.

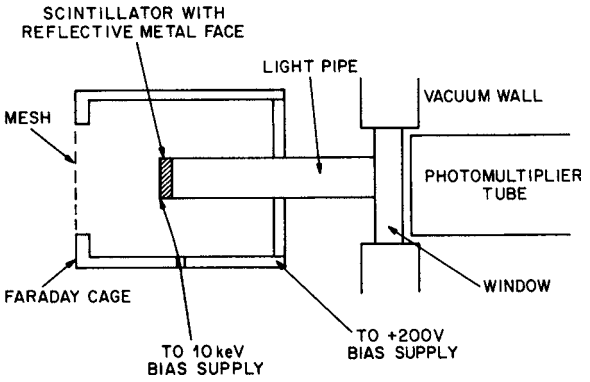


FIGURE 4 The Everhart–Thornley secondary electron detector.

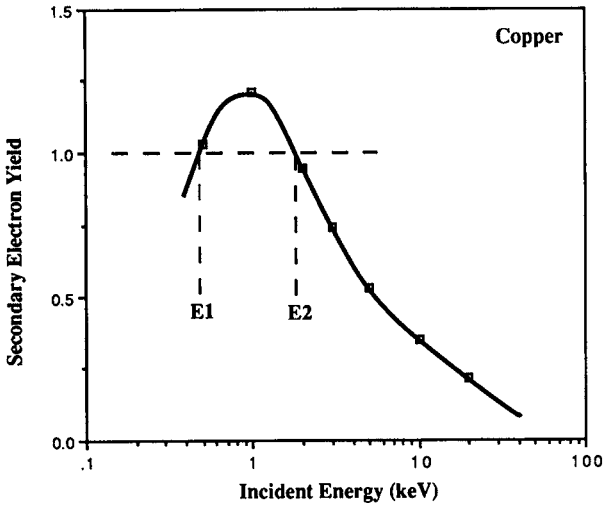


FIGURE 5 Yield of secondary electrons versus beam energy.

will be examined, and their contrast content and other relevant properties will be described.

A. Secondary Electrons

The most popular signal in use in the SEM is that due to secondary electrons. These are electrons, with energies ranging from 0 to 50 eV with a most probable value of about 4 eV. Because of this low energy, secondaries may readily be collected by using a positively biased detector. In the usual arrangement (Fig. 4), described in its original form by Everhart and Thornley, the detector consists of a scintillator which emits light under the impact of electrons. The light travels along a light pipe, through a vacuum window, and into a photomultiplier. Because the light output from a scintillator varies linearly with the energy of the arriving electron the low energy secondary electrons are accelerated to an energy of 10 keV by means of a biasing potential applied to the front face of the scintillator. To prevent this voltage from deflecting the beam the scintillator is shielded by a Faraday cage, made of wide-spacing mesh, maintained at about +200 eV. The field produced by this bias on the cage, typically about 100 V/cm at the sample, is sufficient to ensure that 50% or more of the secondaries emitted from the specimen are collected. In the low pressure SEM a variety of methods are used to collect the secondary electrons based upon their ionization of the atmosphere.

The yield of secondary electrons, for normal incidence of the beam, varies with the atomic number of the sample and with the incident beam energy. Typically at 10 keV the yield is about 0.1 secondaries per incident electron, but this value rises rapidly as the beam energy falls (Fig. 5), reaching a value of unity, for metals and semiconductors, for some beam energy E_2 typically in the range 0.5 to 3 keV, reaching a maximum value of 1.1 to 2 and then falling as the energy is further reduced. For dielectrics and insulator, E_2 is usually lower, 0.2 to 1.5 keV.

If the specimen is not electrically conductive, then if examined at an energy $E > E_2$, more electrons are deposited in the sample than can be emitted, and the specimen charges negatively. If examined at an energy $E < E_2$, then the sample emits more electrons than it received and so charges positively. But if the sample can be examined at the energy $E = E_2$, then a stable image can be obtained without the need to make the specimen conductive. In some cases, as for example when X-ray microanalysis is to be performed, it is necessary to observe a poorly conducting sample at a high beam energy, which is where the low pressure SEM has a unique advantage. The presence of the residual gas can neutralize the effect of the charging, although unlike the energy balancing approach, it does not eliminate it. Otherwise, in the conventional SEM, it is necessary to uniformly coat the surface with 2–10 nm of carbon, or a metal such as Au-Pd or Cr.

In principle, the magnitude of the secondary yield at beam energies $E > E_2$ is a function of the chemistry of the surface since the secondary yield varies slowly with atomic number. However, in most current SEMs the vacuum is sufficiently poor that the surface is always covered with many monolayers of contamination, such as hydrocarbons, deposited from the pumping system. Because the secondary electrons can escape only when produced within a few nanometers of the surface, changes due to the surface chemistry are therefore usually masked.

The dominant imaging mechanism for the secondary electron signal is topographic contrast. This arises because an increase in the angle of incidence θ between the beam and surface normal will generate a greater number of secondaries lying within escape depth from the surface and consequently an increase, as about secant (θ), in the number leaving (Fig. 6). Changes in surface relief will

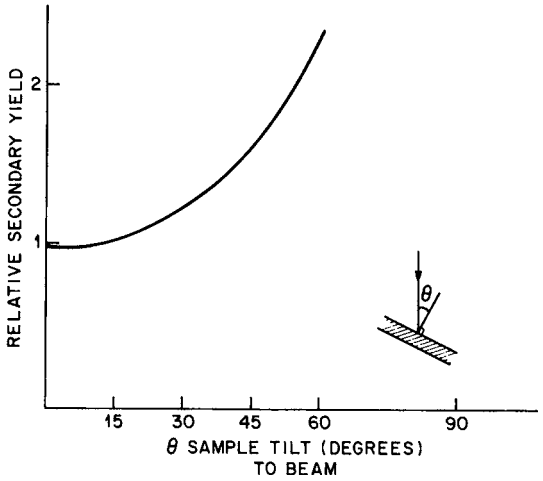


FIGURE 6 Yield of secondary electrons versus tilt.

therefore lead to a signal which rises and falls as the effective angle of incidence varies. This effect can be further enhanced by tilting the sample to the beam. This will both increase the average signal and increase the change in signal resulting from a given change in the angle of incidence, so improving the contrast. Secondaries from areas facing toward the detector will be collected with somewhat higher efficiency than those from faces pointing away from the detector. The resulting picture (Fig. 7) resembles a "real world" image in which the sample, as viewed from the direction of the electron source, is illuminated by a diffuse light source placed at the detector. The intuitive ease with which such an image can be interpreted results from the correspondence between the form of the variation of the secondary yield with incidence angle and Lambert's Cosine Law for the reflection of light from an inclined surface. This, coupled with the three-dimensional quality

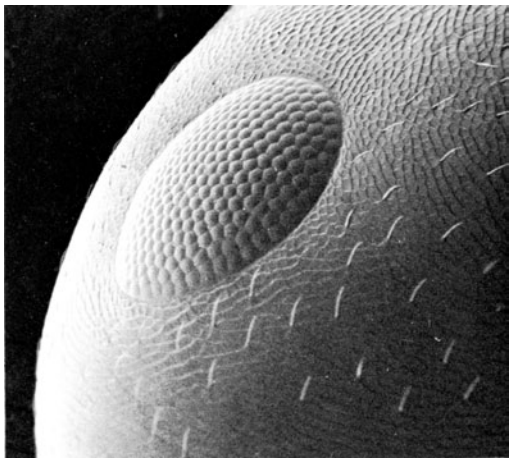


FIGURE 7 Secondary electron image of ant's head and eye.

which comes from the high depth of field, does much to explain the popularity of both the SEM and the secondary electron image.

Secondaries are produced by both the incident primary, and the exiting backscattered, electrons. The ratio of these two components varies with the material of the sample, but typically only 20 to 30% of the secondaries detected are generated by the primary electrons. These SE1 secondaries are those which carry the high spatial resolution image information, since they are produced within the escape depth from the surface and within 2 to 5 nm of the incident probe axis. The secondaries, SE2, generated by the backscattered electrons emerge from an area whose diameter is comparable in size with the electron range in the solid, which may be many microns, so they carry no high-resolution information. Because these electrons are in the majority they reduce the signal contrast carried by the SE1 electrons. From Eq. (7) this implies that the incident beam current must be increased to achieve an adequate signal-to-noise ratio, and this in turn means from Eq. (4) that a larger probe diameter, and so worsened spatial resolution, must be used. Secondary electron imaging is thus ultimately limited by signal-to-noise constraints as much as by fundamental considerations. The best modern instruments, using field emission guns, can achieve resolutions in the range 1–2 nm, while typical commercial SEMs can display 2–5 nm on suitable specimens.

Secondary electron imaging has found wide application in all areas where a wide magnification range, high spatial resolution, and great depth of field is useful. Examples include the study of whole cells, the investigation of the microstructure of alloys, and the study of fracture surfaces in metals, ceramics, and woods. In these, and other similar situations, the special properties of the secondary electron image and the SEM complement the abilities of the conventional optical microscope. One technique of particular value has been that of "stereoimaging" in which two images of the same area are recorded at different sample tilts relative to the incident beam direction. When these two images are viewed through a suitable optical device which presents them separately to the left and right eyes, the parallax produced by the tilt is interpreted by the brain to produce a single "three-dimensional" picture. Detailed quantitative measurements of quantities such as surface roughness, and cavity size, can be obtained from such stereoimages, which have therefore found important application in areas such as tribology and fatigue.

1. Voltage Contrast

Secondary electrons are collected by the field produced by the potential on the Faraday cage surrounding the detector.

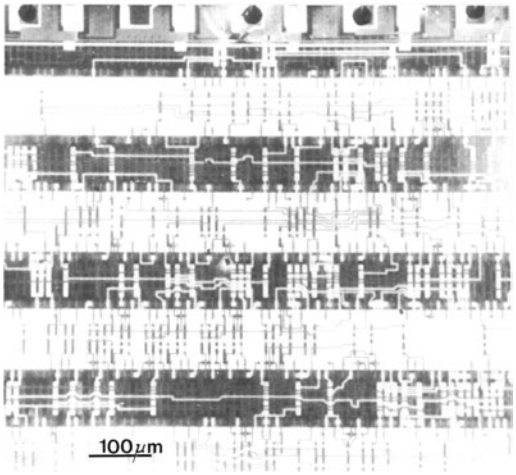


FIGURE 8 Voltage contrast image of integrated circuit.

If different portions of the sample surface are at different potentials with respect to ground, then the potential difference between these areas and the detector cage will also vary and so the collection field will change. In addition, areas that are negative will repel the secondary electrons they emit, while regions that are positive will collect some fraction of the secondaries that they produce. The net result of these effects is that the SE signal is dependent on the surface potential, with negative areas appearing bright and positive areas appearing dark (Fig. 8). This technique of "voltage contrast" is of major importance in the semiconductor industry, because it provides a method for investigating the potentials of small regions, such as the micron-wide conductor lines that interconnect components in an integrated circuit chip. By the use of circuitry designed to linearize the variation of collection efficiency with surface potential, the electron beam can be used to measure surface potentials with a precision of a few tens of millivolts. The SEM thus functions as a voltmeter, but does not require a mechanical probe to select the region to be measured. This means that micron-scale areas can be probed, and it also has the benefit that the measurement does not itself have to draw any current from the circuit under test.

These advantages can be further enhanced by actually operating the circuit of interest under normal conditions while observing it in the SEM. The voltages at various points in the circuit may be changing rapidly with time, but they can still be measured by repetitively switching the incident beam on and off at a frequency which is related to the frequency of the voltage at the point to be observed. For example, if the potential was varying sinusoidally at a rate of 10 kHz (i.e., with a period of 100 μs), then if the beam were switched on for 10 μs at intervals of 100 μs the potential at the point of interest would always be the same when examined, and it could thus be measured in the usual way. The time varying signal has therefore been "frozen" by a technique which is analogous to the stroboscopic imaging technique used optically. By maintaining the sampling rate at the same value, but shifting the time at which the beam is turned on, the potential at some other portion of the cycle could be observed, allowing the complete shape of the voltage waveform at the chosen point of interest in the circuit to be determined. The SEM can thus also be operated like a sampling oscilloscope. Observations can be made in this way on analog and digital circuits operating at frequencies up to several hundred megahertz.

2. Magnetic Contrast

Contrast can also be obtained in the secondary electron mode from the magnetic leakage fields that exist close to the surface of uniaxial magnetic materials such as cobalt, or the field produced by magnetic devices such as recording heads or recording tape. Secondary electrons leaving the sample travel through these fields and so are deflected by the Lorentz force. Depending on the direction of the magnetic field vector and the initial direction of the electron this force can either increase or decrease the probability that the electron will be collected by the detector. Figure 9 shows how this mechanism allows the magnetic domain structure of a cobalt single crystal to be seen. Because both the large scale internal domains and the smaller surface closure domains produce a leakage field, the full details of the domain configuration are visible. This type of contrast depends on the saturation flux of the material, but typically generates high (10–20%) contrast levels making it an easy technique to apply, although the sample must be correctly oriented with respect to the detector for the contrast to be visible.

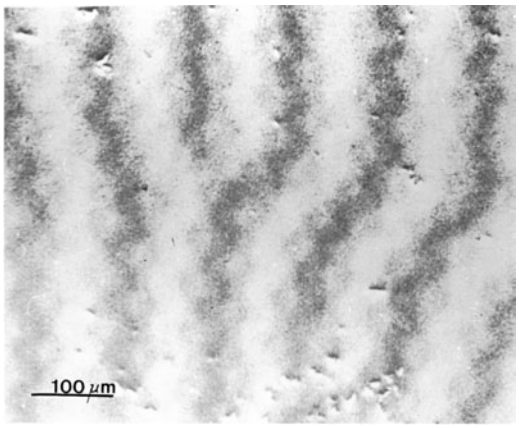


FIGURE 9 Image of magnetic domains in cobalt single crystal.

The SEM technique offers a resolution that is effectively limited only by the scale of the magnetic structure itself, and also has the merit that it is ideally suited for dynamic studies such as the observation of domain structure variations with temperature, stress, or applied magnetic field. The only practical limitation is usually that of preparing samples with a sufficiently flat surface to eliminate topographic contributions to the magnetic image information.

B. Backscattered Electrons

Backscattered electrons are those which leave the specimens with energies between about 50 eV and the incident beam energy. These electrons emerge from depths of up to about 0.3 to the electron range R, where R can be approximated as

$$R = 0.0276AE^{1.67}/(Z^{0.889}\rho)\,\mu\text{m}, \qquad (8)$$

where A is the atomic weight of the sample (g/mol), Z is the atomic number, ρ is the density (g/cm^3), and E is the beam energy in keV. For medium atomic weight elements, and beam energies in the range 10 to 20 keV, the range is of the order of a few microns. The backscattered image therefore contains information from regions beneath the sample surface, but the high energies also enable these electrons to be conventionally collected in the LPSEM.

The yield η of backscattered electrons is independent of the beam energy (over the range 1 to 100 keV) but is a monotonic function of the atomic number Z of the specimen. η can be approximated by the function

$$\eta = -0.0254 + 0.016Z - 0.000186Z^2. \qquad (9)$$

For compounds and alloys Z can be replaced by the arithmetic mean value of Z derived from the chemical composition. η varies between 0.05 for carbon ($Z = 6$) to about 0.6 for gold ($Z = 79$), thus at typical SEM beam energies the backscattered signal is larger than the secondary electron signal. However, because the backscattered electrons are relatively high in energy they are not easily deflected toward a detector. Thus efficient collection of this signal requires a detector which subtends a large solid angle at the specimen. Typically this is achieved by using an annular p–n junction or Schottky barrier solid state device directly above the sample, and concentric with the beam. With such an arrangement 50% or more of the backscattered signal can be collected.

1. Atomic Number Contrast

The variation of η with Z produces image contrast which is directly related to the mean atomic number of the irradiated volume of the specimen. Thus in materials such as multiphase alloys, regions of the specimen with different atomic numbers will display contrast, the magnitude of which will depend on the relative change in Z (Fig. 10). Under normal imaging conditions regions with an effective difference of only about 0.5 units in Z can be distinguished. The fact that the variation of η with Z is not only monotonic but almost linear also makes it possible to perform a simple form of chemical microanalysis on a sample by measuring the variation in the backscattered signal and comparing this with the signal produced under identical conditions from suitable pure element standards.

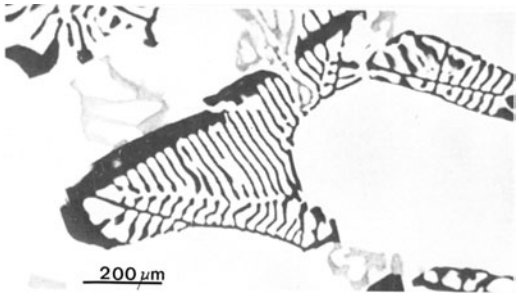


FIGURE 10 Atomic number contrast from aluminum–copper alloy.

The atomic number contrast effect is of importance in many biological applications, since heavy metal reagents having affinities for specific groups can be used as stains. In the backscattered image these stained regions then appear bright against the predominantly low atomic number carbon matrix. Because the backscattered electrons are collected from depths up to about 0.3 of the electron range [as determined from Eq. (8)] the labeled regions can be observed at significant depths beneath the surface of a specimen. For example, at 15 keV a penetration in excess of 1 μm is possible, although the lateral spatial resolution will, correspondingly, be poor.

The atomic number contrast is superimposed on any topographic contrast present from the sample, since changes in surface orientation also lead to changes in the backscatter yield. In addition, the fact that the backscatter electrons travel in straight trajectories from the sample to the detector produces shadowing of any surface relief. The topographic and chemical components of the signal can be separated, at least partially, by using multiple detectors. An annular detector divided into two halves will minimize topographic contrast and maximize atomic number contrast when the signals are added, since the shadowing seen by one segment of the detector will in general not be present in the signal from the other segment, whereas the situation will be reversed if the two signals are subtracted since both segments will see the change in signal due to the change in atomic number in the same way.

2. Electron Channeling Patterns

Electron channeling contrast arises directly from the interaction of the electron beam with a crystalline sample. For beams incident at random angles on a crystal there is an approximately constant probability of the electron being scattered out of the sample and being collected. However, if the electron is incident along one of the symmetry directions of the crystal lattice then the electron may penetrate, or channel, a significant distance into the specimen before being scattered. In this case the chance of the electron emerging and being collected is reduced. The backscattered signal from a crystal recorded as a function of the angle of incidence of the beam therefore shows a variation which displays directly the symmetry elements of the lattice. Figure 11 shows an example recorded from the (111) face of a silicon crystal. The 3-fold symmetry associated with the (111) face is immediately evidence, in the arrangement of bands which cross at the (111) pole. The angular width of the bands is twice the Bragg angle θ_B where

$$\theta_B = \lambda/2d. \tag{10}$$

λ is the electron wavelength and d is the lattice spacing. For electrons of 20 keV energy λ is 0.087 Å, so lattice spacings of a few angstroms will produce bands with a width of the order of 0.02 rad, about 1°. Higher order reflections produce families of lines parallel to the band at spacings of θ_B. By calibrating the angular width of the display from a known crystal, the symmetry and lattice spacings of an unknown crystal can rapidly be determined. The pattern observed will not change when the sample is moved laterally because the symmetry will not alter, but tilting or rotating the crystal will change the symmetry and cause the channeling pattern to change as if rigidly fixed to the lattice. This fact can be used to build up a channeling map of a crystal, showing all the symmetry elements exhibited by the lattice.

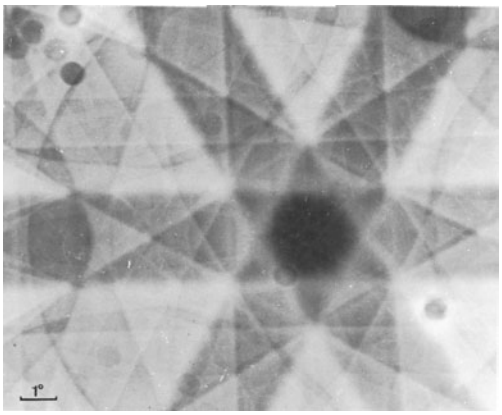


FIGURE 11 Electron channeling pattern from the (111) face of a silicon crystal.

Channeling contrast comes from the top few hundred angstroms of the crystal surface, the quality of the pattern is therefore very dependent on the quality of the surface. Small amounts of surface contamination or mechanical damage will lead to the degradation, or elimination, of the pattern. The electron channeling pattern can therefore be used as a sensitive test of surface condition and crystal quality, and electron channeling has been widely used in the study of wear and deformation, and in the investigation of rapid thermal annealing by laser or electron sources. Similarly, information can be learned at higher resolution from a static beam when it is referred to as an electron backscattered pattern. However, the signal strength is dramatically reduced necessitating either extremely long exposures or expensive photosensitive detectors to record the patterns.

3. Magnetic Contrast

Contrast from the magnetic domain structure of specimens that have no external leakage field (i.e., materials with cubic magnetization) can be obtained in the backscattered mode. The contrast arises from the Lorentz deflection of the beam as it travels through the magnetic induction inside the sample. If the sample is inclined to the beam then domains of opposite magnetization will deflect the beam slightly closer to, or further away from, the surface so modifying the backscatter yield and producing an image in which the domains show as bright or dark. The contrast is very small, typically only 1% or less, so high beam currents and large probe sizes are required. If the sample is not inclined to the beam then differential deflection of the beam occurs only at domain walls, which then show up as dark or bright lines in the image. This effect is extremely weak, producing contrast levels of 0.3% or less, and has been observed only on materials with high saturation magnetization.

C. Electron Beam–Induced Currents

In one important additional mode of operation the specimen itself is used as a detector. When the incident electron beam enters a semiconductor, or insulator, it creates electron–hole pairs, promoting electrons across the band gap into the conduction band leaving behind holes in the previously filled valence band. Typically the energy e_{eh} required to create one electron–hole pair is about three times the band gap, thus in silicon e_{eh} is about 3.6 eV. A single 10 keV incident electron could therefore generate nearly 3000 such carrier pairs. Under normal conditions the electrons and holes will recombine within a

short time (microseconds or less); however, if an electric field is imposed on the material then the electrons and holes will drift in opposite directions and the resultant current flow can be detected through an external circuit. The electron beam has thus produced localized conductivity in the semiconductor. When the electric field is produced by an external source this effect is known as beta conductivity. In the vicinity of a p–n junction, however, there exists a space-charge depleted region and an associated field; a current detector connected to the two sides of the junction will record the current due to the motion of the beam induced carriers when the electron beam is at, or close to, the junction. When the beam moves away from the junction the carriers must first diffuse back to the depletion region before they are separated by the field, and the signal will therefore fall at a rate determined by the minority carrier diffusion length. This electron beam–induced current will have a peak amplitude I_{cc} given by

$$I_{cc} = I_b E / e_{eh}, \tag{11}$$

where E is the beam energy and I_b is the incident current. I_{cc} may therefore be several thousand times the incident current.

The electron beam–induced current (EBIC) mode of operation provides a powerful tool for the examination of both semiconductor devices and materials. Figure 12 shows the EBIC image obtained by connecting an amplifier across the positive and negative power lines of an integrated circuit. Whenever the electron beam comes close to either of the junctions in each of the transistors making up the circuit the resultant current flow will be detected, and the image thus displays all of the active devices in the circuit. By varying the beam energy, junctions at significant depths beneath the surface can be observed, permitting a three-dimensional plan of the device to be derived. Observations on semiconducting materials may be made by forming a Schottky barrier, such as gold, onto the surface. This produces a depletion field extending for several microns beneath the full extent of the barrier region. An amplifier connected between the barrier and the material will therefore detect the EBIC signal generated by the beam. Electrically active regions in the material, such as dislocations, stacking faults, and grain boundaries, which lie within the electron range from the surface will cause locally enhanced recombination rates for the carriers and so lead to a fall in the collected current. Defects therefore appear as dark lines in the image. The EBIC mode thus provides a direct technique for the observation of potentially damaging electrically active defects in material prior to processing. Figure 13 shows the EBIC image of dislocations in polycrystalline silicon used in solar cell manufacture.

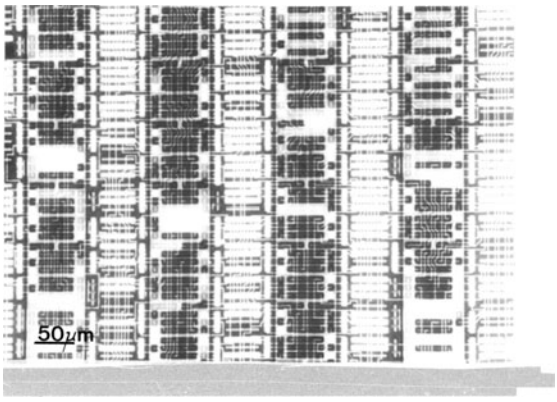


FIGURE 12 Electron beam–induced current image of integrated circuit.

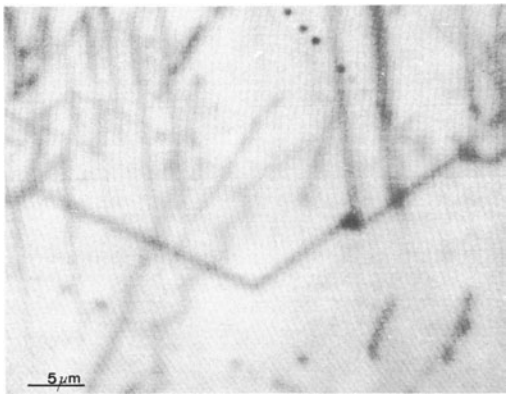


FIGURE 13 Electron beam–induced current image of single dislocation in silicon.

IV. OTHER IMPORTANT INTERACTIONS

A. Fluorescent X Rays

Energetic electrons can transfer some of their energy to atoms along their trajectory by ionizing inner shell electrons. The resultant excited state of the atom can be relaxed by the production of either a fluorescent X ray or an Auger electron which can carry away the excess energy. Since the energy, in either case, is directly related to the binding energy of the inner shell that was excited, detection and measurement of either radiation will permit a chemical identification of the element that was ionized. The usual practice is to detect the fluorescent photon by means of a high efficiency solid-state energy dispersive X-ray spectrometer placed a few millimeters away from the sample. Such a device in conjunction with a pulse height analyzer (multichannel analyzer) provides a visual display of the energy spectrum of all X rays leaving the sample. A rapid,

quantitative elemental microanalysis can therefore be obtained from an area of the sample selected by the position of the beam. In this mode the SEM is acting as an electron microprobe.

The spatial resolution of the X-ray microanalysis is determined by the beam interaction volume. X rays can be produced by all electrons with an energy greater than the binding energy of the inner shell being observed, so then, as an electron travels through a material it can generate X rays along its trajectory until its energy drops below the critical value required. Depending on the initial energy of the electron and the critical excitation energy of the characteristic line X rays may therefore be generated from a volume several microns in diameter even for a point source.

The combination of chemical microanalysis, on a micron scale, with the versatile imaging abilities of the SEM has found wide application in all areas of the science, to the extent that 70% of all new SEMs are delivered with an attached X-ray system. Conventional systems permit the detection and identification of all elements in the periodic table from sodium upward, with a trace sensitivity of the order of 1%. More advanced systems utilizing wavelength discrimination extend this level of sensitivity for the light elements by orders of magnitude which can be particularly significant. In addition to the simple identification of elements, spectral data can be reduced to give quantitative chemical composition data by comparison with known standard elements and compounds coupled with reduction schemes which correct for factors such as absorption of X rays in the sample and variations in the detector efficiency with photon energy.

B. Cathodoluminescence

The recombination of electrons and holes mentioned earlier (Section III.C) results in the release of energy, some of which may be radiated as light in the visible or infrared portions of the spectrum. This cathodoluminescence may be collected by any light-sensitive device. However, since the yield of photons is very small (about 1 per 10^6 incident electrons) the signal is weak and carefully optimized optics must be used. Typically an ellipsoidal mirror is employed, with the sample placed at one of the foci and the detector (such as a photomultiplier) at the other. In this way a high fraction of all of the emitted radiation can be collected. When a spectral analysis is required the detector is replaced by a light-guide which directs the light to the entrance slit of a suitable spectrometer.

The spatial resolution of the cathodoluminescence signal is set by the beam interaction volume, and is therefore of the order of the beam range [Eq. (9)]. It can, in principle, be made as small as desired by using a sufficiently low beam energy and is not subject to the usual optical diffraction limit because no optical imaging is involved. However, the weakness of the signal requires that a high-intensity beam, and thus a large probe size, be used. In practice the achievable resolution is of the order of 1 μm.

The magnitude and spectral distribution of the cathodoluminescence signal depend on several factors. The spectrum, for a semiconductor, will show a peak at the band gap energy, and this peak will shift and broaden as the sample temperature is increased. Spectral measurements are therefore usually made with the sample held at liquid nitrogen temperatures or lower (100 K or less). The mechanism which produces cathodoluminescent radiation is in competition with other nonradiative modes or recombination, such as those involving deep traps. Since radiative recombination is enhanced by the presence of impurities the magnitude of the light monotonically follows the doping level over a wide range of variation. The presence of electrically active defects, which act as local centers for nonradiative recombination, will result in a fall in the light output, so that dislocations and other defects will show as dark features in the image such as those visible in Fig. 14 which shows strain induced defects in GaAs. The image information is thus comparable with that obtained in EBIC imaging, although in this case no junctions or Schottky barriers are required.

Cathodoluminescence is also observable from many biological and geological materials which have natural properties as scintillators. In such cases the spectrum is more complex than that from a semiconductor, major spectral peaks usually being identifiable with the

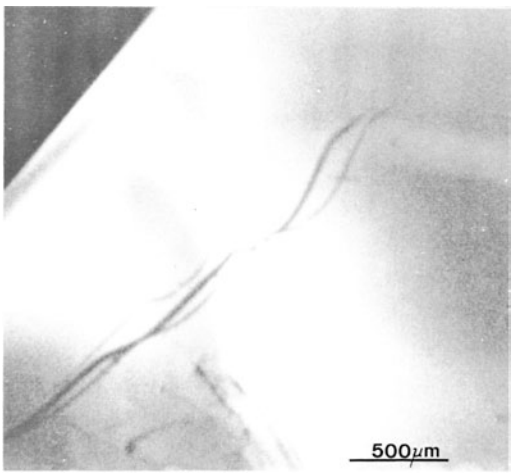


FIGURE 14 Cathodoluminescence image of defects in a gallium arsenide wafer.

excitation of interatomic bonds in the specimen. Potentially cathodoluminescence offers a suitable tool for the detailed examination of such bonds; however, the ionizing action of the electron beam may, in many cases, lead to a destruction of the bonds before the spectrum can be recorded.

C. Thermal Wave Microscopy

The end product of the interaction of the electron beam with the specimen is heat, and this too can be used to examine the specimen. If the electron beam is switched on and off at a steady rate, usually 50 to 100 kHz, then the periodic heating and cooling of the specimen (of the order of $\pm 1^{\circ}$) around the beam impact point will lead to the production of an elastic stress wave propagating away from the beam point. This elastic wave travels freely through the specimen and can readily be detected by bonding a piezoelectric transducer at any convenient point to the sample. The output from this device can then be used to form an image in the normal way. Because the efficiency with which the thermal energy is converted into mechanical energy is low, usually 10^{-8} or so, incident beam currents of the order of 1 μA or more at the sample are required and the spatial resolution attainable is consequently limited by the probe size to a few microns.

The amplitude of the signal detected will depend on many factors, those which determine the magnitude of the temperature rise around the beam impact, and those which determine the efficiency with which this thermal energy is converted into mechanical energy. Contrast in the thermal wave image therefore contains information about the thermal properties (specific heat, conductivity, density) and the elastic moduli (expansion coefficients). In addition the elastic wave may be scattered by gross mechanical defects such as cracks and interfaces encountered within the specimen. Thermal wave images therefore contain a complex mixture of information. Typical examples of the application of the method have included the visualization of subsurface cracks in integrated circuits, the determination of doping profiles, and studies of elastic and plastic deformation in solids.

V. CONCLUSION

The scanning electron microscope is by far the most widely used electron-beam instrument. This popularity results from the unrivaled versatility of the instrument's imaging modes, its relative ease of operation, its modest cost, and a level of performance which places it conveniently between the optical microscope and the more complex transmission electron microscope. The SEM has proved to be of especial value in the semiconductor industry, since it provides the only feasible method for examining high density, submicron feature devices. However, while one in every two SEMs is sold to the electronics industry, the instruments in use by biologists, geologists, and metallurgists have proved equally productive, as evidenced by the rapidly growing literature in each of these fields. The Bibliography points to some of the important sources providing more detailed information on SEM based research in these, and other, areas.

SEE ALSO THE FOLLOWING ARTICLES

CHARGED-PARTICLE OPTICS • PARTICLE SIZE ANALYSIS • POSITRON MICROSCOPY • SCANNING PROBE MICROSCOPY • TRANSMISSION ELECTRON MICROSCOPY • X-RAY ANALYSIS • X-RAY PHOTOELECTRON SPECTROSCOPY

BIBLIOGRAPHY

Goldstein, J. I. *et al.* (1981). "Scanning Electron Microscopy and X-ray Microanalysis," Plenum Press, New York.

Hawkes, P. W. (ed.) (1985). "The Early Days of Electron Microscopy," (Advances in Electronics and Electron-Optics Supplement 16). Academic Press, London.

Heinrich, K. F. J. (1981). "Electron Probe Microanalysis," Van Nostrand-Reinhold, New York.

Holt, D. B., and Joy, D. C. (1989). "SEM Characterization of Semiconductors," (Techniques of Physics, Vol. 12). Academic Press, London.

Newbury, D. E. *et al.* (1986). "Advanced Scanning Electron Microscopy and X-ray Microanalysis," Plenum Press, New York.

Oatley, C. W. (1982). The history of the scanning electron microscope. *J. Appl. Phys.* **53,** R1–R13.

Speckle Interferometry

Harold A. McAlister
Georgia State University

GLOSSARY

Adaptive optics The real-time correction of atmospheric distortion through closed-loop detection and warping of an optical surface in a telescope system in a manner opposite to that produced by turbulence.

Airy pattern Diffraction pattern produced when light from a point source passed through a circular aperture shows a bright central peak surrounded by a series of decreasingly bright concentric rings. The central bright peak is known as the Airy disk.

Aperture Diameter of the mirror or lens that first collects light in a telescope and forms an image of the object under observation.

Coherence Property possessed by two or more beams of light when their fluctuations are highly correlated.

Diffraction Change in direction of a ray of light when passing an obstacle or a change that occurs when passing through some aperture, because of the wave nature of light.

Dispersion Differential bending of light of different wavelengths by a refracting medium. Atmospheric dispersion occurs when light from a star is observed at angles other than directly overhead.

Interference Combination of two or more coherent beams of light that leads to the reinforcement of intensity where wave crests overlap and the cancellation of intensity at locations where wave crests overlap with wave troughs.

Isoplanatic angle Angular size of a single cell of atmospheric coherence, determined by the physical size of the cell and its elevation.

Photoelectric effect Emission of electrons from certain materials that occurs when light of a wavelength less than some critical value strikes the material.

Pixel Picture element or the smallest resolution element within an image. It may be a light-sensitive grain in a photographic emulsion or a small rectangular element in an electronic detector.

Rayleigh limit Theoretical diffraction limit to angular resolution that occurs when the bright central peak of the Airy pattern of a point source is located in the first dark ring of the Airy pattern from a second point source. This limit, in radians, is given

by 1.22l/D where l is the wavelength and D is the aperture.

Seeing Quality related to the blurring of star images by atmospheric turbulence usually expressed in the angular size of a blurred image.

Troposphere Lowest region of Earth's atmosphere extending to an elevation of about 11 km and in which weather phenomena occur.

Wave front Surface that is at all points equidistant from the source of light and from which light rays are directed in a perpendicular manner.

Wavelength Distance between successive wave crests or any other repetition in phase in a beam of light. Visible wavelengths are from about 400 to 650 nm; infrared wavelengths begin beyond 650 nm.

GALILEO first used the telescope for astronomical purposes nearly four centuries ago, and the subsequent history of astronomy is closely tied to the quest for telescopes of ever-increasing power. The fundamental capability of a telescope is determined by its aperture. The aperture area determines the amount of light that a telescope can collect and hence sets an effective limit to the faintest observable objects. Light-gathering power has been the primary motivation behind the construction of large telescopes as astronomers seek to understand the most distant and faintest components of the universe. The largest telescopes now in existence can collect millions of times the amount of light in comparison with the human eye; and new generations of telescopes with 8- to 10-m apertures incorporate technologies that may be extrapolated to 50 m.

The second capability tied to aperture is angular resolution, the detection of fine structural detail in images. Whereas impressive gains in light-gathering power have been achieved since Galileo, blurring produced by Earth's atmosphere has set a limit to resolution equivalent to that achievable by an aperture of only about 10 in. Not until 1970 was a method discovered that has allowed astronomers to attain the full theoretical resolution of large-aperture telescopes. This method is known as speckle interferometry. Speckle techniques are providing a wealth of information for wide classes of astronomical objects and have become the standard method for measuring orbitals motions in resolved binary star systems.

I. ASTRONOMICAL SEEING AND RESOLUTION

Earth's atmosphere places serious limitations upon astronomical observations for two primary reasons. First, the atmosphere is not equally transparent at all wavelengths and certain wavelength regimes are completely inaccessible from the ground. Observations from telescopes orbiting above the atmosphere have to a certain extent circumvented this obstacle. Second, turbulence within the atmosphere produces image blurring that seriously degrades the ability of telescopes to resolve detail in images. Beginning in the late 19th century, astronomers realized the importance of locating observatories at sites with exceptionally stable air in order to achieve the best possible seeing conditions. It is fortunate that the properties that lend favorable astronomical seeing are also consistent with transparency, and modern observatories are typically located at relatively high elevations in very dry climates.

The intrinsic limiting ability of a telescope to resolve fine angular detail is set by the diffraction properties of light. For a telescope such as the 4-m-aperture Mayall reflector on Kitt Peak in southern Arizona observing at a wavelength of 550 nm, the center of the visible region of the spectrum, the Rayleigh limit is approximately 0.035 arcsec, an angle equivalent to that subtended by a nickel seen from a distance of 75 miles.

Unfortunately, the atmosphere thwarts the realization of such resolution and imposes an effective limiting resolution from 1 to 2 arcsec, a degradation in resolution by a factor of roughly 50. Some locations on Earth offer seeing that is occasionally as good as 0.2–0.3 arcsec, but even these rare and superb seeing conditions are an order of magnitude worse than what would be obtained under ideal circumstances. One obvious option is to put telescopes into orbit above the atmosphere. Indeed, a primary justification for the 2.5-m Hubble Space Telescope (HST) has been its ability completely to avoid atmospheric blurring. For the foreseeable future, ground-based telescopes will continue to be built with apertures significantly larger than their far more expensive space-borne counterparts. The two 10-m Keck telescopes on Mauna Kea in Hawaii as well as the new 8-m-class telescopes in the northern and southern hemispheres offer three to four times higher resolution than HST if the challenge of atmospheric blurring can be overcome. Thus, extensive resources have been expended to exploit methods for correcting turbulent blurring by special imaging techniques and especially through adaptive optics.

In 1970, the French astronomer and optical physicist Antoine Labeyrie pointed out an elegantly simple method for circumventing atmospheric seeing conditions to achieve diffraction-limited resolution. Labeyrie's method, which he named speckle interferometry, takes advantage of the detailed manner in which the blurring occurs in order to cancel out the seeing-induced effects.

The atmosphere is a turbulent medium with scales of turbulence ranging from perhaps hundreds of meters down to turbulent eddies as small as a few centimeters. The

turbulence arises from the dynamics of the atmosphere as driven by Earth's rotation and the absorption of solar radiant energy, which is converted into the thermal energy content of the atmosphere. Turbulence along does not induce "bad seeing," rather it is the variation in density from one turbulence region to another that causes rays of light to be refracted from otherwise straight paths to the telescope.

Light from a star spreads out in all directions to fill a spherical volume of space. The distances to stars are so great that a typical star can be envisioned as a point source illuminating a spherical surface on which the telescope is located. Because the radius of this imaginary sphere is so enormously large in comparison with the telescope aperture, the light entering the telescope at any instant can be pictured as a series of parallel and plane wave fronts. Equivalently, all rays of light from the star to the telescope can be considered as parallel rays perpendicular to the incoming wave fronts. To this simple picture we must add the effects of the atmosphere.

A telescope accepts the light from a star passing through a cylindrical column of air pointing to the source and having a diameter equal to the telescope's aperture. If the column of air were perfectly uniform, the incoming wave fronts would remain flat, and a perfect Airy pattern could be formed. The density variations accompanying turbulence exist at elevations throughout the troposphere. The cumulative effect of these fluctuations can be modeled as being equivalent to patches across the telescope entrance aperture, or pupil, such that within one such patch, rays of light remain roughly parallel (or alternately, the wave front remains nearly flat). A given patch, or coherence cell, will produce some net tilt of the parallel bundle of rays and will retard the entire bundle by some amount referred to as a piston error. A telescope whose aperture is stopped down to match the diameter of these cells, a quantity commonly referred to as r_0, would produce an instantaneous image in the form of an Airy pattern. From one instant to the next, a given r_0-size cell moves because of winds and will even dissolve on a slightly longer time scale because of the dynamics of turbulence.

A telescope with an aperture larger than r_0 will at any time contain $(D/r_0)^2$ coherence cells, each of which produces some random tilt and piston deviations on its bundle of rays from the mean of these deviations. At any instant, there will be some fraction of these deviations arising from points distributed randomly throughout the aperture that have nearly identical tilt and piston errors. The light from these coherent subapertures undergoes interference to produce a fringe pattern that shows regions of brightness and darkness. These bright regions are called speckles, and each speckle is in essence a distorted or noisy version of an Airy pattern. The entire distribution of speckles at any instant fills a region whose size corresponds to the Airy disk of a single r_0-size aperture. Under typical seeing conditions, the coherence cell size r_0 is from 8 to 15 cm, with seeing conditions degrading as r_0 becomes smaller. When $r_0 = 15$ cm, the seeing disk diameter will be about 1 arcsec. The parameter r_0 improves with wavelength to the exponent 6/5, making turbulent blurring a nonissue at infrared wavelengths longer than 10 μm. The twinkling of starlight as observed by the unaided eye is produced by the rapid passage of individual coherence cells, each of which is significantly larger than the pupil of the eye, across the line of sight with the resultant apparent rapid motion of the star arising from the random tilts from each successive cell. Planets do not appear to twinkle because their disks are sufficiently extended in angular size to average out the tilts from a number of cells at any instant.

The rapid motion and dissolution of seeing cells requires the use of short exposure times in order to record a speckle pattern. For exposures longer than the atmospheric redistribution time t_0, speckle patterns will blur into the classic long-exposure image of a star in which the image profile intensity drops off in a Gaussian-like manner to fill the arcsecond-scale seeing disk. Experience has shown that exposure times no longer than about 0.01 sec are typically sufficient to freeze the speckles. Atmospheric conditions vary considerably from place to place and from time to time, and values of t_0 less than 0.001 sec have been encountered. Exposures on such a short time scale will permit the detection of so few photons, even from a bright object with a large telescope, that speckles cannot effectively be recorded. Fortunately, such rapid seeing conditions are rare.

II. SPECKLE CAMERA REQUIREMENTS AND TECHNOLOGY

The first step in Labeyrie's method of speckle interferometry is to record the speckle images from large telescopes using specially designed cameras. Several factors tend to reduce the amount of light available to a speckle camera. Because speckles result from the interference of overlapping wave fronts of light, their sharpness or contrast is related to the size of the wavelength region over which the interference occurs. The production of useful speckle requires that the recorded wavelengths be restricted to ranges of no more than a few tens of nanometers, the remaining light transmitted by the telescope being rejected by filters that transmit only in some preselected wavelength region. Therefore, as much as 95% of the otherwise available light from the object must be filtered out prior to recording speckles. The necessity for exposure times shorter than t_0 provides a weak level of illumination for speckle imagery in comparison with classical astronomical imagery, where

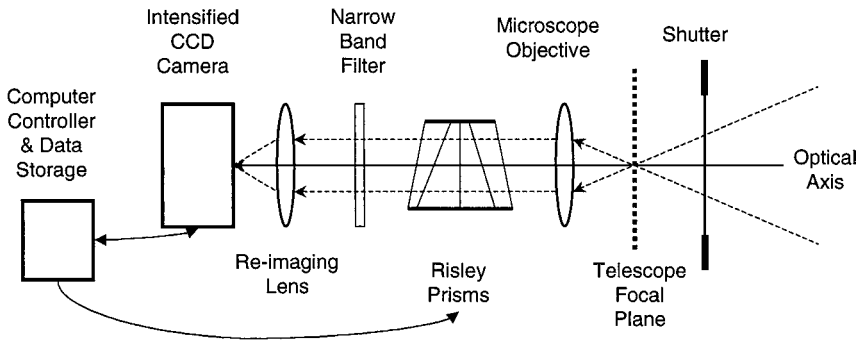


FIGURE 1 Schematic view of a simple speckle camera system.

exposures of many minutes or even hours are commonly used to integrate light. Finally, because each speckle must be resolved by the speckle camera system, a very high magnification is required in order to sample each speckle with more than a single pixel on the camera detector. The high magnification leaves few photons available per pixel during a sub-t_0 exposure time. Clearly, speckle cameras must employ very sensitive detectors in order to record speckle patterns from any but the very brightest stars.

A speckle camera system is schematically illustrated in Fig. 1. Light from the telescope comes to a focus just in front of a microscope objective that serves the purpose of increasing the effective magnification of the telescope by a factor of perhaps 20–30. The beam from the microscope objective is very slowly converging or can be made collimated to be brought to a focus later by a lens. An interference filter provides the required spectral filtering. For telescope apertures larger than about 1.5 m, the Airy disk size is small enough so that spectral dispersion from the atmosphere produces noticeably elongated speckles, even over the rather narrow spectral regions used. Risley prisms provide a useful means for introducing dispersion that can be adjusted to the appropriate amount in the direction opposite to that arising from the atmosphere to cancel out this effect. The magnified, spectrally filtered, and dispersion-compensated beam then passes through a shutter before striking the detector.

The first generation of speckle cameras used film systems coupled to magnetically or electrostatically focused image intensifier tubes. Electronic gating of these intensifiers provided an effective means of shuttering to freeze atmospheric turbulence. Current speckle cameras typically incorporate intensified CCD arrays operating at standard video frame rates. Although CCDs have quantum efficiencies much higher than photographic emulsions, the readout noise is typically high in comparison with the low photon rates inherent in the speckle recording process. Thus, some degree of image intensification is still required. Other types of digital detectors are capable of "photon counting," i.e., providing a continuous list of photon arrival coordinates and arrival times. This offers advantages to some applications of speckle imaging.

A speckle picture of a bright star is shown in Fig. 2. Had Labeyrie's idea occurred fully to the previous generation of astronomers who had only photographic emulsions at their disposal, they would have been frustrated by the lack of technology that would enable speckle interferometry to be carried out beyond a demonstration stage.

III. APPLICATION TO HIGH-RESOLUTION IMAGING

Labeyrie showed mathematically that the production of speckles by a telescope and the atmosphere is mathematically equivalent to the convolution of the point-spread function (PSF) of the atmosphere with the object intensity

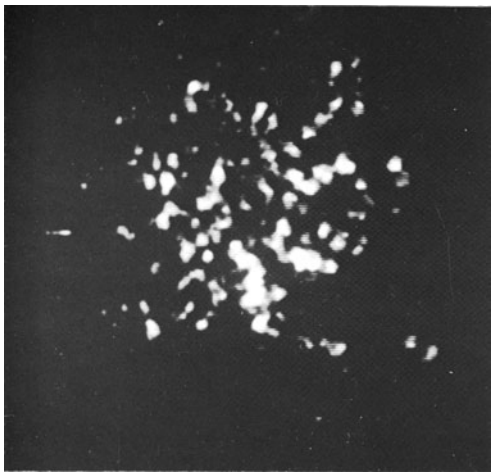


FIGURE 2 A speckle image of a single bright star showing discrete speckles, each of which is a representation of an Airy disk corresponding to the 4-m aperture of the telescope at which this picture was taken.

distribution on the sky. The so-called convolution theorem states that the Fourier transform of a convolution is equal to the multiplicative product of the individual Fourier transforms of the two quantities to be convolved. This fact is the essence of speckle interferometry. If, for example, speckle images are obtained of one of the handful of the highly evolved and rare supergiant stars that are near enough to the Sun so that their angular diameters are larger than the Airy disk of a telescope, then the convolution theorem shows that the star diameter can be deconvolved from the speckle data by dividing the Fourier transform of speckle images of the supergiant star by the transform of speckle images from an unresolved star to cancel out the effects of the atmospheric PSF.

Whereas nature provides only a few supergiant stars resolvable by speckle methods with the largest existing telescopes, there are thousands of binary star systems suited to the technique. The two stars comprising a binary system are bound in orbit around a common center of mass by their mutual gravity and may be so close together or the system may be so far from the Sun that the angular separation of the components is smaller than the seeing disk or even smaller than the Airy disk. For binaries with angular separations in the regime of 0.03–0.3 arcsec, speckle interferometry currently provides the best method for accurately measuring their orbital motions. Such measurements lead to the determination of stellar masses, quantities that are relatively rarely known and yet play a vital role in our theoretical understanding of the origin and evolution of stars. In fact, the mass essentially predetermines the entire course of evolution of a star.

The two speckle patterns arising from two stars in a binary system are highly correlated as long as the light from the two stars passes through the same collection of coherence cells, a condition known as isoplanicity and illustrated in Fig. 3. The two stars then give rise to speckle patterns that are very nearly identical but are displaced from each other by an amount equal to the angular separation of the binary. The isoplanatic angle is typically a few arcseconds, so that binary stars with angular separations less than this amount will produce speckle patterns with very high point-to-point correlation. Figure 4 is a speckle image of a widely separated binary star in which the high degree of correlation between the speckle patterns from the two stars in the system is obvious. For systems closer in separation than that in Fig. 4, the two speckle patterns will merge together and overlap in such a way that every speckle will be doubled, an example of which is shown in Fig. 5. Thus, a speckle image of a binary star produced at a 4-m-aperture telescope will contain hundreds of individual representations of the binary star geometry.

A simple method of analysis of binary star speckle data is provided by the method of vector autocorrelation. A vector autocorrelogram (VAC) can be produced by plotting all the pairings among speckles that occur within a speckle image. Because the pairing corresponding to the actual binary star pair occurs very frequently compared with every other random pairing among any two speckles in an image, the binary star geometry stands out in

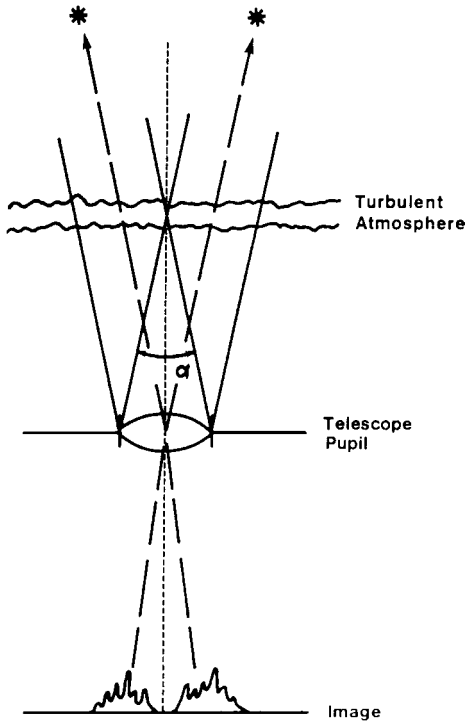


FIGURE 3 The condition of isoplanicity for a pair of stars separated by the angle at which the separate beams of light are just no longer passing through any common atmospheric turbulence. The speckle patterns in this case would be uncorrelated or nonisoplanatic.

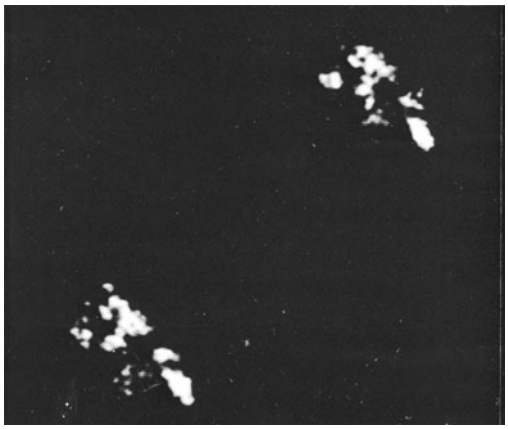


FIGURE 4 A 2-arcsec-separation binary showing a very high correlation between the separate speckle patterns of the two stars indicative of isoplanicity.

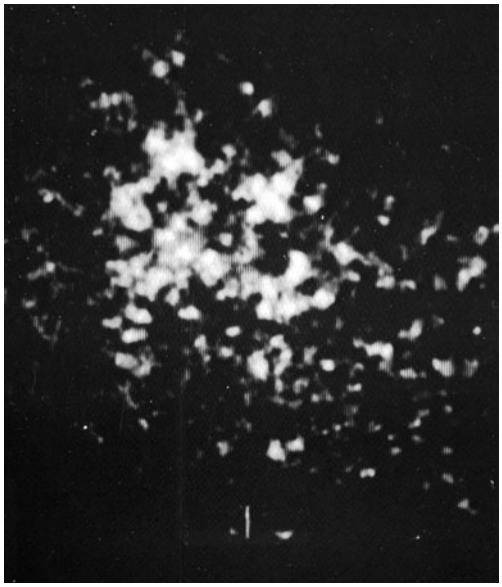


FIGURE 5 A speckle photograph of a binary star having an angular separation of 0.25 arcsec showing an apparently single speckle pattern caused by the complete overlap of the patterns from the two stars.

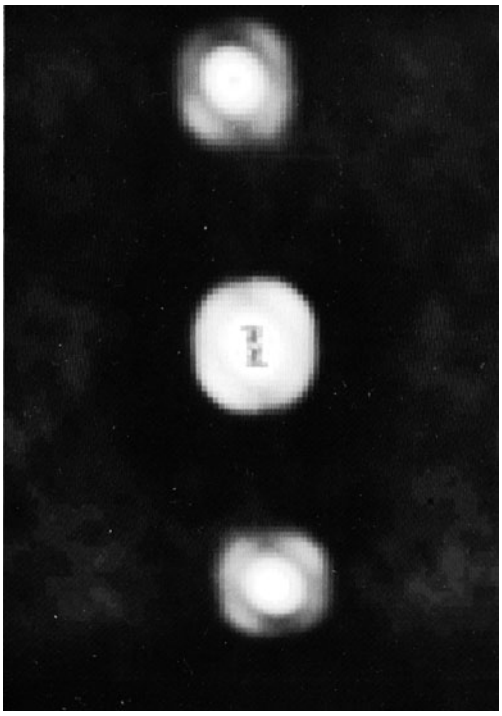


FIGURE 6 The computer-generated vector autocorrelogram of approximately 1800 speckle images of the same system as shown in Fig. 5 showing the characteristic central peak accompanied by identical peaks arising from the binary star geometry.

a VAC. The individual VACs from many speckle images of a binary can be added together to further increase the relative contribution from the binary star pairing. A VAC possesses a bright central peak because of the pairing of every speckle with itself. This central peak is accompanied by two identical peaks on opposite sides, one of which is produced by the sum of all speckle pairings of star 1 with respect to correlated speckles from star 2, while the other results from pairings of star 2 with respect to star 1. This simple algorithm can operate sufficiently rapidly in a computer so that data can be reduced in real time as they are produced by a digital speckle camera system. A vector autocorrelogram of approximately 1800 speckle images of the same binary as is exemplified in Fig. 5 is shown in Fig. 6. The binary peaks in a VAC are superimposed upon a background that gradually slopes away from the central peak. This background results from all the uncorrelated speckle pairings and has a radius determined by seeing. It can be subtracted using a VAC produced under the same seeing conditions but for a single star.

Deconvolution methods, such as the one described previously, provide an effective means for studying objects that have simple geometries, such as single stars whose diameters are resolved or binary star systems describable only by the angular spacing of the two components and their relative orientation. However, for more complex objects, it is necessary that actual images of these objects with diffraction-limited detail be reconstructed from the speckle data. A fundamental problem in such reconstructions is that the atmosphere randomly distorts the relative phase of the wave front to such an extent that the real phase information is lost in a speckle image. To produce a real image requires the incorporation of these missing phases and the amplitudes of the incoming waves. Algorithms have been developed that work with phases or amplitudes, typically in an iterative approach, to attempt image reconstruction, and a number of interesting examples show the promise of these methods.

IV. ASTRONOMICAL RESULTS FROM SPECKLE INTERFEROMETRY

A. Stellar Angular Diameters

The largest existing telescopes have apertures that just begin to yield resolutions capable of resolving the diameters of stars, and even so, only in the case of the nearest supergiants. Thus, speckle interferometry has been able to measure such stars as Betelgeuse in the constellation Orion and the half dozen or so other supergiants closest to the Sun. Speckle image reconstructions for Betelgeuse have shown this star, whose surface would extend beyond the orbit of Mars if Betelgeuse were to replace the Sun,

to be nonuniformly luminous and to be surrounded by a very extended shell of gas and dust. Although such diameter measures are rare, an important result is that speckle measures have been consistent with results obtained by a variety of other techniques.

B. Binary Stars

As a class, binary stars are ideally suited to speckle interferometry. Suitable candidates for measurement exist in almost unlimited supply, and speckle methods are providing hundreds of new discoveries while measuring previously known systems with greatly enhanced accuracy over classical methods. There currently exist more than 30,000 speckle measurements of some 4000 binary star systems, many of which have been measured through sufficient orbital motions to permit calculation of the parameters describing these motions. Approximately 10% of these systems had not been previously resolved. These orbital elements lead to the determination of the masses of the components, quantities that can then be used to confirm or improved theoretical models of stellar structure, formation, and evolution. In addition to its high-resolution capability, speckle interferometry allows the measurement of known binary stars with an order-of-magnitude increase in accuracy in comparison with classical techniques. This ultimately leads to an increase in the accuracy of mass determinations.

Because of the increased resolution of speckle interferometry over other methods, speckle surveys for new binaries are able to penetrate into separation regimes not previously detectable. These surveys, although rather limited in extent because of the strong competition for time on the largest telescopes, are supporting already existing evidence that the majority of stars in our galaxy exist in binary or multiple star systems of higher complexity.

C. Infrared Speckle Interferometry

A particularly interesting area for speckle interferometry and image restoration has been the application of these methods at infrared (IR) wavelengths, where the atmosphere is more benign than at optical wavelengths and where there exist classes of objects of moderate complexity that are ideal candidates for high-resolution imaging. As wavelengths increase, both t_0 and r_0 increase, and the observational requirements are relaxed. The first IR speckle observations were made with single-pixel detectors across which were scanned at high speeds the images of objects to be analyzed. This approach has been used to measure the heated dust shells surrounding such supergiants as Betelgeuse and Antares and other hot and highly evolved stars. The sizes of protostellar objects from which normal stars will eventually evolve have been measured. These objects are typically enshrouded in dense dust clouds that obscure the visible radiation while radiating at IR wavelengths because of heating from the central hot-star-forming gas. IR sources discovered by standard methods have been found to be highly complex, and very faint and cool companions have been found in orbit around a number of stars. The star T Tauri, the prototype of a class of stars thought to represent the transition between protostars and normal hydrogen-burning stars, has been found to have a companion. This provides a rare opportunity to study the circumstances surrounding the formation of a binary star system.

IR speckle methods were particulary advanced by the advent of extremely sensitive solid-state detectors with full two-dimensional pixel coverage. These powerful new devices combined with the wealth of objects to which they can be applied make IR speckle interferometry and direct imaging an extremely productive tool for exploring a variety of phenomena, especially those associated with young stars and star-forming regions.

Surveys for companions to young and pre-main sequence stars in several star-forming regions have shown that the occurrence of duplicity for these young objects is at least as high as for older stars like the Sun. This indicates that the formation of binary and multiple star systems is a natural consequence of the earliest stages of star formation.

V. THE FUTURE OF HIGH-RESOLUTION ASTRONOMY

The successes of speckle interferometry have firmly established high-resolution astronomy as an important scientific enterprise. Users of new, large telescopes are paying special attention to their interferometric applications as well as to their light-collecting potential. Whereas not long ago it was considered necessary to go into space to achieve significant gains in resolution, there are now numerous plans to further extend the boundaries of resolution from the ground using single giant telescopes or arrays of telescopes.

The 1990s saw the inauguration of several large telescopes with apertures in the 8- to 10-m range. The European Southern Observatory has built four 8-m telescopes on Cerro Paranal in the Chilean Andes in a facility known as the Very Large Telescope (VLT). The VLT has the light-gathering power of a single 16-m telescope. A consortium of institutions led by the University of Arizona is building the Large Binocular Telescope (LBT) on Mt. Graham, Arizona, consisting of two co-mounted 8.4-m light-collecting mirrors. The VLT, LBT, and the twin

10-m Keck telescopes represent a great advance in our ability to collect light from the faintest objects in the universe. They also push resolution through speckle imaging and adaptive optics applied to each of their large individual apertures.

Many problems in astrophysics call for resolutions comparable to apertures of 100 m and larger. Individual telescopes with such enormous dimensions are not likely to be built in the foreseeable future, but the prospects for very high resolution imaging are being realized through the construction of synthetic large apertures using methods analogous to those employed at radio wavelengths.

Arrays of telescopes, each of which may have a relatively small aperture, can be distributed along the ground to effectively synthesize an aperture hundreds of meters across. Optical wavelengths are far more challenging to this approach than are the much longer radio wavelengths, and the technology suited to multiple-telescope optical arrays matured significantly in the 1980s and 1990s. Interferometer arrays have been built in Australia, Europe and the United States. The VLT and twin Keck telescopes are both being equipped for "long baseline" interferometry by linking the large telescopes together with smaller "outrigger" telescopes. Experience from speckle interferometry with single telescopes has gone far to improve our knowledge of how the atmosphere will affect such arrays as we strive for a gain of a factor of 100 over the resolution now provided by speckle methods. Several dedicated optical/interferometric arrays now in operation are expected to provide important new data pertaining to stellar physics as well as the first images of the surfaces of stars of a variety of masses, diameters, and temperatures. These facilities are likely to be the progenitors for a next-generation array incorporating dozens of telescopes with apertures of 4–8 m synthesizing a kilometer-size aperture.

Space-borne interferometers should be operational this decade and are expected to provide unprecedented increases in our ability to directly measure distances to stars and to add to the list of known extrasolar planetary systems. Ultimately, space interferometry may yield the first images of the surfaces of planets around other stars.

High-angular-resolution astronomy is providing a revolutionary approach to viewing the universe. This is remarkable progress in the three decades since Labeyrie invented speckle interferometry.

SEE ALSO THE FOLLOWING ARTICLE

ATMOSPHERIC TURBULENCE • BINARY STARS • IMAGING THROUGH THE ATMOSPHERE • INFRARED ASTRONOMY • OPTICAL INTERFEROMETRY • RADIO-ASTRONOMY INTERFEROMETRY • STELLAR STRUCTURE AND EVOLUTION • TELESCOPES, OPTICAL

BIBLIOGRAPHY

Bates, R. H. T. (1982). "Astronomical speckle imaging," *Phys. Rep*. **90,** 203.

Dainty, J. C. (1984). "Stellar speckle interferometry," *Topics Appl. Phys*. **9,** 255.

Fischer, D. (1996). "Optical interferometry: Breaking the barriers," *Sky Telescope* **92,** 36–43.

Labeyrie, A. (1976). "High-resolution techniques in optical astronomy," *Prog. Opt*. **14,** 47.

Labeyrie, A. (1978). "High-resolution techniques in optical astronomy," *Annu. Rev. Astron. Astrophys* **16,** 77.

Lena, P., Lebrun, F., and Mignard, F. (1998). "Observational Astrophysics," 2nd ed., Springer, Berlin.

McAlister, H. A. (1985). "High angular resolution measurements of stellar properties," *Annu. Rev. Astron. Astrophys*. **23,** 59.

McAlister, H. A. (1988). "Seeing stars with speckle interferometry," *Am. Sci*. **76**(2), 166–173.

McAlister, H. A. (1996). "Twenty years of seeing double," *Sky Telescope* **92,** 28–35.

Stable Isotopes as Tracers of Global Cycles

Robert T. Gregory
Southern Methodist University

GLOSSARY

Asthenosphere The portion of the Earth's mantle below the lithosphere, 10s to 100s of kilometers below the surface, that is actively convecting at rates of centimeters per year and deforming in the solid state. Its temperature is close to the melting point so that it is presumed to be the source region for basalt, the most voluminous product of mantle melting.

Atmosphere The gaseous envelope of the Earth that is predominantly a mixture of nitrogen and oxygen. Water vapor and carbon dioxide are important greenhouse gas constituents of the atmosphere.

Continental crust The portion of the Earth's surface that normally stands above mean planetary radius (sealevel) and rises abruptly from abyssal sea depths (−4 km) to include shallow marine platforms and large land masses. Its composition is not the result of direct melt extraction from the mantle.

Hydrosphere The water portion of the Earth's surface including the oceans, water vapor in the atmosphere, fresh surface water, and groundwater.

Igneous and metamorphic processes The major processes that control the formation and evolution of the crust of the earth. Igneous processes involve the melting and solidification of rocks. Metamorphic processes involve the transformation of rocks chemically and texturally by temperature and pressure.

Lithosphere The rigid outer rock layer of the Earth consisting of the crust plus uppermost mantle. It is a thermal and mechanical boundary layer for the Earth separating cold brittle surface rocks, whose thermal and mechanical properties change rapidly with depth, from hot mantle rocks whose physical properties change slowly with depth.

Low temperature geologic processes Processes that occur at near Earth surface conditions down to several 10s of kilometers of burial beneath the surface. Isotopic compositions of earth material exhibits larger ranges under these conditions.

Magmatic arcs Belts of igneous rocks associated with the upper plate of a subduction zone, a boundary where large tectonic plates collide with one plate returning to the mantle.

Mantle of the Earth The major silicate portion of the Earth left over from iron core extraction that extends from the core, 3000 km below the surface, to within on average 10–35 km of the surface.

Paleoclimatology The study of ancient climates that makes use of evidence from the paleogeographic distribution of the depositional environments of sedimentary rocks and their chemical and isotopic composition as well as the fossil record of plants and animals.

Plate tectonics The paradigm that divides the Earth into about a dozen lithospheric plates that either diverge at midocean ridges, converge at subduction zones or slip past each other. These plate boundaries explain the age of distribution of the oceanic crust, its heat flow and its bathymetry.

THE LIGHT STABLE isotopes of hydrogen, carbon, nitrogen, oxygen, and sulfur (HCNOS) are natural tracers of geologic/planetary processes. These elements and their isotopes are major constituents of common compounds that are found in gas, liquid, and solid form. As such, they make excellent tracers of interactions between major reservoirs such as the atmosphere, hydrosphere, lithosphere, and asthenosphere of a planetary body. Their isotopic ratios are readily measured by mass spectrometric methods.

I. THERMODYNAMIC BASIS

As light stable isotopes with significant relative mass differences (e.g., 2:1 for D and H or 18:16 for oxygen isotopes), there are measurable differences in the physical properties of end member light stable isotopic compounds (e.g., D_2O and H_2O or $H_2{}^{16}O$ and $H_2{}^{18}O$). These differences in physical properties manifest themselves in their thermodynamic properties so that the ratios of isotopically substituted compounds represent chemical activity products (α's) that have temperature significance. Exchange reactions are the special class of chemical reactions where the only difference between the reactants and the products is in the isotopically substituted species. The compounds are otherwise identical. In general, the fractionation factor α is the ratio of isotopes of two substances, Phases 1 and 2:

$$\alpha_{12} \equiv \frac{R_1}{R_2}, \tag{1}$$

where the measured isotopic ratio of D/H, $^{13}C/^{12}C$, $^{18}O/^{16}O$, $^{34}S/^{32}S$ represents the R values in the ratio and the subscript 12 refers to the ratio of Phase 1 relative to Phase 2. The equilibrium fractionation factor is related to the equilibrium constant of the exchange reaction by

$$K_{12} = (\alpha_{12})^n, \tag{2}$$

where n refers to the number of exchangeable sites in the exchange reaction. For example, the isotopic exchange reaction between water (w) and calcite (cc) is

$$3H_2{}^{18}O + CaC^{16}O_3 \Leftrightarrow 3H_2{}^{16}O + CaC^{18}O_3 \tag{3}$$

It has a fractionation factor given by

$$\alpha_{CC-W} = \frac{\left(^{18}O/_{16}O\right)_{calcite}}{\left(^{18}O/_{16}O\right)_{water}} \tag{4}$$

For the calcite–water exchange reaction, important for paleoclimatology, the equilibrium constant is α^3. In terms of practical laboratory measurements, it is convenient to use the delta notation whereby measured ratios are referenced to an isotopic standard.

$$\delta \equiv \left[\frac{R_{sample}}{R_{standard}} - 1\right] * 10^3. \tag{5}$$

Table I gives the standards and some approximate ranges for important reservoirs of hydrogen, carbon, oxygen, and sulfur. The ranges are approximate because the rate of new data acquisition is growing exponentially with a more than doubling of the mass spectrometers in use over the last decade.

The delta notation is useful because α's are typically $1 \pm$ a number $\ll 1$. In the delta notation, fractionation factors become

$$\alpha_{12} = \frac{1000 + \delta_1}{1000 + \delta_2}. \tag{6}$$

A useful relationship that takes advantage of the identity, $\ln(1 \pm \varepsilon) \approx \pm\varepsilon$ for $\varepsilon \ll 1$, is given by

$$\Delta_{12}(T) \equiv \delta_1 - \delta_2 \approx 1000 \ln\alpha_{12}. \tag{7}$$

At equilibrium, the fractionation factor is related to the temperature of exchange. Measured fractionations (differences in delta values) are therefore proportional to temperature of exchange. Harold Urey's recognition of these properties led to the application of stable isotopic measurements to fossils containing carbonate to infer the temperature of the ancient oceans.

As in any multiphase system with a series of reactions that characterize the interactions, one additional constraint is necessary to specify the system. In the case of

TABLE I Standard and Some Approximate Ranges for Important Reservoirs of Hydrogen, Carbon, Oxygen, and Sulfur

Isotopic system		Standard	
Important reservoirs	**δ values**	**Important reservoirs**	**δ values**
Oxygen isotopes: $^{18}O/^{16}O$		Standard mean ocean water	
Crust and mantle		Surface reservoirs	
Mantle	$\delta^{18}O \approx 5.5$	Marine carbonate	$+17 < \delta^{18}O < +34$
Moon	≈ 5.5	Chert	$+14 < \delta^{18}O < +40$
Midocean ridge basalt	≈ 5.7	Clastic rocks	$+8 < \delta^{18}O < +30$
		Hydrosphere	≈ -1
Altered oceanic crust	$+1 < \delta^{18}O < +16$	Ocean	≈ 0
Eclogite	$+1 < \delta^{18}O < +8$	Meteoric water	Average ≈ -4
Hornblende biotite granitic rocks	$+6 < \delta^{18}O < +10$	Ice caps	Average ≈ -35
Two mica granitic rocks	$+10 < \delta^{18}O < +14$		
Carbon isotopes: $^{13}C/^{12}C$		Pee Dee Belemnite	
Crust and mantle		Surface reservoirs	
MORB CO_2	-6 ± 2	Marine carbonates	$-2 < \delta^{13}C < +4$
Igneous rock (C)	$-33 < \delta^{13}C < -7$	Authigenic carbonates	$-30 < \delta^{13}C < +30$
Graphite	$-38 < \delta^{13}C < -5$	Organic matter	Average $\approx -26 \pm 10$
Diamond	peak -5, tail -35	Carbon dioxide	-7
		Methane	-47 peak, tail -110
Sulfur: $^{34}S/^{32}S$		Canyon Diablo Troilite	
Crust and mantle		Surface reservoirs	
Meteorite	$\delta^{34}S \approx 0$	Seawater	$+20$
Mantle/MORB source	≈ 0	Evaporites	$+12 < \delta^{34}S < +35$
		Sedimentary sulfides	$-40 < \delta^{34}S < 20$
Igneous rocks	$-9 < \delta^{34}S < +19$		—
Ore deposit sulfides	$10 < \delta^{34}S < +10$		
Extreme sulfides	-35; $+25$		
Hydrogen: D/H		Standard mean ocean water	
Crust and mantle		Surface reservoirs	
Mantle/MORB	$\approx -80 \pm 5$	Ocean	$\delta D \approx 0$
Mantle phlogopites	$-85 < \delta D < -50$	Meteoric water	≈ -22
Mantle amphiboles	$-90 < \delta D < -30$	Ice caps	≈ -270
Igneous rocks	$-85 < \delta D < -40$	Hydrosphere	≈ -10
Seafloor serpentine	$-60 < \delta D < -30$	Marine sediments	-65 ± 20

paleotemperature measurements, Urey assumed that seawater was a large reservoir of oxygen and that changes in temperature would be reflected in changes in the isotopic composition of carbonate fossils and not in the original seawater. This illustrates the additional constraint on stable isotopic systems, i.e., conservation of isotopes.

The molar volumes of isotopically substituted compounds are virtually indistinguishable for most common compounds so that typical light stable isotopic exchange reactions involving solid materials are insensitive to changes in pressure normally encountered under lithospheric conditions making them suitable for geothermometry (the study of temperature within the Earth). Potentially, stable isotopic differences between phases reflect temperatures at the time the ratio is locked in under equilibrium conditions. This typically occurs when some chemical reaction occurs because stable isotopic exchange kinetics are very sluggish (joule-scale driving forces instead of kilojoule driving forces for chemical reactions).

The mass balance constraint allows the inference of exchange between interacting reservoirs of different bulk isotopic composition and in particular the movement of material between the various high and low temperature reservoirs in the Earth and between gas, fluid, and solid reservoirs.

II. STABLE ISOTOPES IN GLOBAL CYCLES

A. The Rock Cycle: Emphasis on Oxygen Isotopes

In general, low temperature geologic processes impart large isotopic heterogeneity in surface reservoirs (up to 10s of parts per thousand). In contrast, isotopic interactions during high temperature igneous and metamorphic processes result in small fractionations (parts per thousand or less). For example, the mantles of the Earth and the Moon are inferred, on the basis of the oxygen isotope data, to be homogeneous on large scales. The existence of igneous rocks with oxygen isotope ratios higher than values possible from materials derived from the mantle of the Earth shows that sedimentary materials are recycled to great depth in the Earth and remelted to make certain classes of igneous rocks. Similarly, the existence of high- and low-^{18}O eclogite inclusions in basaltic rocks originally melted from the mantle with normal oxygen isotope ratios is spectacular confirmation that surface plates are recycled into the mantle of the Earth. When compared with the isotopically homogeneous lunar lithosphere reflecting small magnitude igneous fractionations, the existence of water on the Earth enables large isotopic heterogeneities to develop as a result of the Earth's dynamic tectonic regime.

1. Oxygen Isotopic Composition of the Mantle

The oxygen isotopic composition of the mantle is inferred to be very uniform on large scales. The evidence for this comes from the analysis of (1) midocean ridge basalts, the most voluminous product of partial melting of the mantle; (2) the analysis of peridotite nodules brought up as inclusions in magmas of various compositions, generally basalt, but also kimberlite; (3) analysis of peridotite massifs exposed in ophiolite complexes or tectonic slivers brought up along convergent margins; and (4) the isotopic composition of the Moon which may have split off from the Earth's mantle.

All of these measurements suggest that the isotopic composition of the Earth's mantle sits at a δ^{18}O value of about 5.5 per mil with seawater defined as the standard at 0 per mil. Midocean ridge basalts, which are produced at rates of 18 km^3/yr, average out to approximately 5.7 per mil. The heterogeneity observed in MORB sources appears to be on the order of tenths of a per mil.

There is some suggestion that some parts of the continental lithosphere, particularly lithosphere that has remained for significant periods of time as the hanging wall of a subduction zone, are enriched in ^{18}O. These enrichments in the upper mantle wedge are the result of infiltration by slab derived water-rich fluids and enable the generation of small volumes of alkalic rocks with elevated primary mantle δ^{18}O values. The classic example of this phenomenon is the potassic volcanic province of Italy.

2. Oxygen Isotopic Composition of the Oceanic Crust

The oceanic crust, on the basis of seismological studies of the seafloor, dredge hauls, drilling, and studies of ophiolites on land, is layered consisting of an upper sedimentary layer, a middle basaltic volcanic layer, and a lower third layer consisting of gabbroic plutonic rocks. On mature oceanic crust, the sedimentary layer consists of differing proportions of biogenic and terrigenous sediments depending upon the position of the oceanic crust with respect to latitude and the productivity of the surface ocean.

The initial oceanic crust accretes at the midocean ridge with an isotopic composition typical for midocean ridge basalt, i.e., 5.7 per mil. This initially pristine crust immediately reacts with seawater because the ridge axis acts like a heat engine driving hydrothermal circulation throughout the solidified oceanic crust. Within a very short period of time (<1 Myr) the igneous portion of the oceanic crust becomes zoned in δ^{18}O with the upper portion of the crust becoming enriched in ^{18}O and the lower portion of the crust becoming depleted in ^{18}O with respect to pristine midocean ridge basalt. These changes in δ^{18}O values result from the temperature dependence of the exchange between basalt and seawater.

The hydrothermal alteration that results from the circulation of seawater through the oceanic crust produces a hydrated ^{18}O-enriched zeolite-to-greenschist facies upper crust and an ^{18}O-depleted amphibolite-to-granulite facies lower crust (Fig. 1). The boundary between the two zones corresponds roughly to the oceanic Layer 2 to Layer 3 boundaries or the contact between sheeted dike complex and gabbro in ophiolite complexes. This zonation occurs because of (1) the geometry of accretion of oceanic crust, (2) the contrast in temperature between the overlying ocean and the hot magma intruded into the ridge system, and (3) the contrast between the oxygen isotopic composition of seawater and the mantle-derived magmas.

Pelagic sediments overlie the igneous rocks of the oceanic crust and are dominated by two inputs: biogenic detritus from the surface ocean (silica and carbonate) and airborne terrigenous materials derived from the continents or from volcanic eruptions in island arcs that settle through the ocean to reside on the seafloor. The biogenic material is precipitated in the water column by various silica and carbonate secreting organisms. In today's oceans, these materials represent a major sink for the dissolved load of rivers carrying the chemical weathering signature from

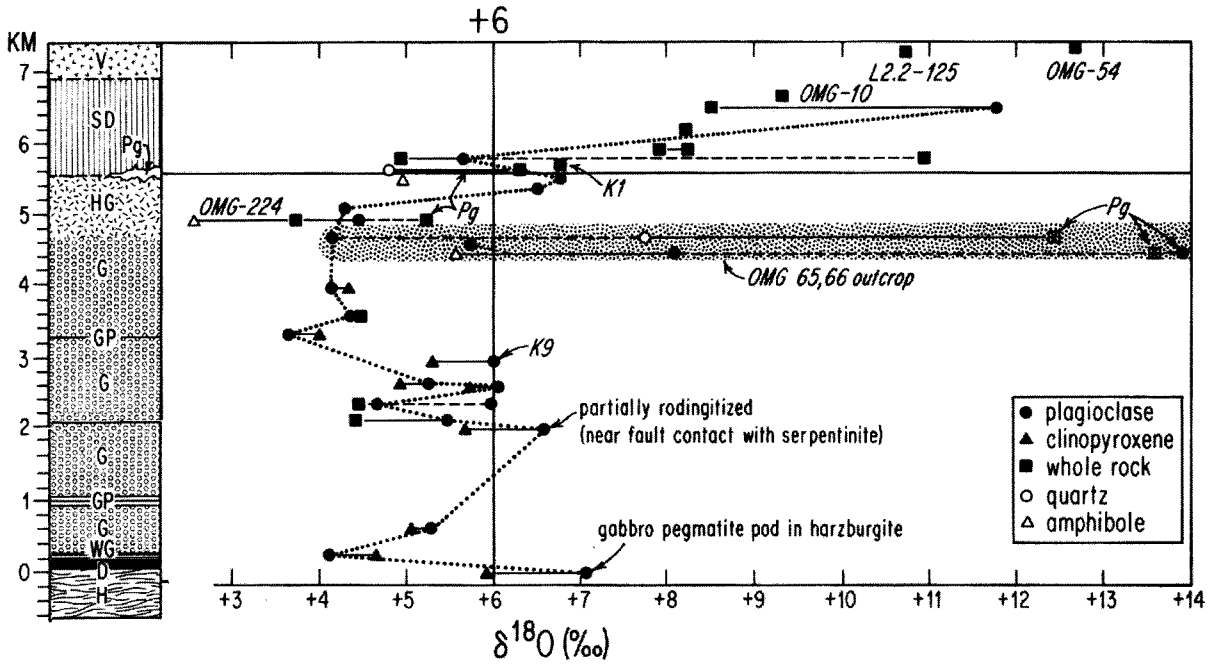


FIGURE 1 A composite structural column through the Samail ophiolite complex shows the typical $\delta^{18}O$ profile through the ophiolite, an analog for oceanic crust. The ^{18}O-enriched upper crust is complementary to ^{18}O-depleted lower crust. Initially, the magmas coming from the asthenosphere under the spreading center are uniform in their $\delta^{18}O$ values, +5.7. Subsolidus exchange with circulating seawater is responsible for the redistribution of ^{18}O in the crust. The balance between the enriched and depleted portions of the crusts indicates that the average bulk fractionation between seawater and mantle is near the steady-state value. Sm–Nd and Rb–Sr determinations exist for the samples noted by their sample numbers. Cross-cutting relationships at the OMG 65, 66 outcrop enable the recovery of most of the temperature history of the hydrothermal system at a single locality.

the continents. The terrigenous component is dominated by clay minerals resulting from continental weathering or chemical weathering of volcanic ash.

Collectively, all of these sediments are enriched in ^{18}O decreasing from siliceous oozes with the highest $\delta^{18}O$ values (>35 per mil), carbonate oozes with intermediate $\delta^{18}O$ values (≈30 per mil), and clays (≈20 per mil depending on their provenance). Because deposition rates for pelagic sediments are very slow, millimeters per thousand years, the high ^{18}O sedimentary layer is relatively thin (less than a few hundred meters).

Taken as a whole, the upper portion of the oceanic crust, Layers 1 and 2, is hydrated, ^{18}O–enriched, enriched in radiogenic isotopes such as ^{87}Sr, and enriched in incompatible elements with respect to the ^{18}O-depleted more refractory and less hydrous gabbroic crust. This structure has implications for the transfer of material from subducted oceanic crust to mantle wedges above subduction zones.

3. Oxygen Isotopic Composition of the Continental Crust

It is very difficult to characterize the average oxygen isotopic composition of the continental crust. This problem can be divided into two parts: platform sedimentary rocks and the underlying basement rocks. The former are typical of miogeoclinal settings consisting of sandstones, shales, and limestone platforms depending upon the paleolatitude. The shales and the carbonate rocks have $\delta^{18}O$ values similar to their pelagic counterparts, although shelf carbonates from shallow, land-locked epicontinental seaways may have been depleted in ^{18}O with respect to modern carbonates.

The $\delta^{18}O$ values of sandstone depend on the provenance of the quartz (Fig. 2). For example, the major quartz components of Archean sandstones such as the Gorge Creek Group in the Pilbara have $\delta^{18}O$ values little shifted away from primary igneous quartz values of approximately 10 per mil. In contrast, turbidite sands from the Paleozoic Lachlan fold belt, eastern Australia or the Ouachita Mountains, south central United States, are derived from both igneous and metamorphic rocks and thus have $\delta^{18}O$ values that are significantly enriched with respect to normal igneous rocks ($\delta^{18}O$ values > +14 per mil). Metamorphic quartz, particularly quartz formed from fluids buffered by crustal sediments, generally has more positive $\delta^{18}O$ values than normal igneous quartz, ranging from 10 to over +20 per mil.

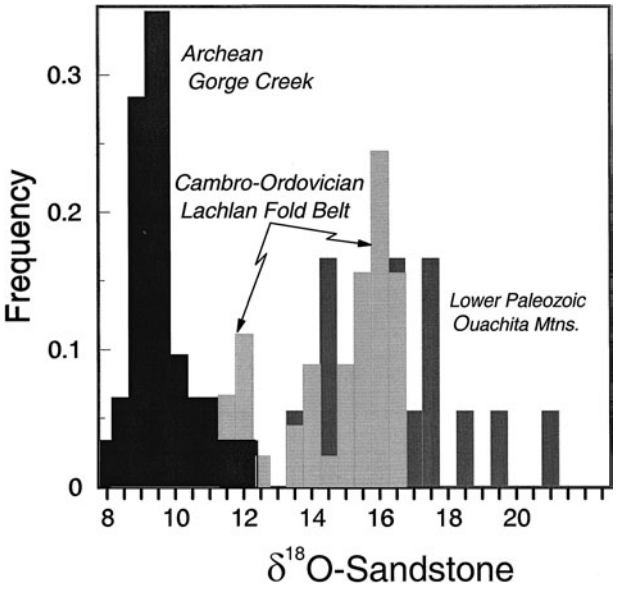


FIGURE 2 $\delta^{18}O$ values from the Archean Earth (Gorge Creek Group, Pilbara Craton, Western Australia) compared with the same from quartzites derived from multiply cycled crust (Lachlan Fold Belt) and the Ouachita Mountains of the south central United states. The Archean (>3 Gyr) Gorge Creek samples exhibit a very narrow range of $\delta^{18}O$ values consistent with the sampling of predominantly igneous quartz. Because of its hardness, quartz persists in the rock cycle as a major component in clastic sediments. Metamorphic quartz veins, formed under crustal conditions, are typically 1–2 per mil more positive in $\delta^{18}O$ than the bulk host wallrocks. As the crust is recycled by plate tectonic processes, the $\delta^{18}O$ values of sandstones increase. Both the Lachlan Fold Belt quartz and the Ouachita Mountain quartz distributions are at least bimodal consisting of a recycled component and a more pristine igneous component.

Underlying the thin veneer of platform sediments is the continental basement that is much more difficult to quantify in terms of its origin and its oxygen isotopic composition. Clearly, the age distribution of the continental rocks suggests some type of continental growth process that occurs at rates of cubic kilometers per year. Plate tectonics provides a mechanism for adding new continental crust at convergent margins.

Magmatic arcs form on the hanging walls of subduction zone and represent new additions to the continental crust. Typically, magmatic arcs are underpinned by voluminous amounts of basalt and capped with composite stratovolcanoes that exhibit the basalt–andesite–dacite series of magmas. In general, the volcanic rocks have $\delta^{18}O$ values slightly enriched from midocean ridge basalt (+6 to +7) reflecting some contribution of slab-derived high-^{18}O material. For the subvolcanic plutonic gabbro–tonalite–granodiorite association, there is evidence of greater involvement of high ^{18}O material originally derived from the surface that has been dragged down into the zone of melt generation and accumulation. As a result the great batholiths of the world have more positive $\delta^{18}O$ values (+8 to +10) than is typical for magmas derived directly from the mantle of the Earth.

A second type of continental basement involves the incorporation of submarine turbidite fans deposited on oceanic crust. Deformation and shortening of the submarine fan by as much as 60% produces a two layer crust consisting of a lower crust of mafic oceanic crustal rocks and an upper crust of chevron folded turbidites (Fig. 3). Intrusion of plutons into this thickened crust marks the stabilization of the new crust which is typically 40 km thick and is in isostatic balance with a surface near sea level. The Lachlan fold belt of eastern Australia represents a good example of this type of crust. The lower crust consists of hydrothermally altered oceanic crust and the upper crust consists of turbidite deposits, alternating sand-rich layers (psammite) and clay-rich layers (pelite). The turbidites carry an ^{18}O signature reflecting the surface history of the rocks weathered away to provide the mass of the turbidite fan. The degree of ^{18}O enrichment depends upon the source rocks and the isotopic composition of the fluids involved in the chemical weathering process.

A third type of continental basement is the granite/greenstone association common in Precambrian terranes. The greenstones typically form dome and basin structures with synclinal keels separating plutons of tonalite–granodiorite–granite association. The greenstones range from undeformed rocks to rocks exhibiting very high strains with a stretching lineation that is predominantly vertical. The greenstone rocks exhibit $\delta^{18}O$ values that are typical of greenschist facies metabasalts throughout geologic history, ≈+9 per mil (Fig. 4).

The granitic rocks have variable $\delta^{18}O$ values depending on their source regions, but typically exhibit $\delta^{18}O$ values of typical normal igneous rocks (+8 to +10 per mil). Two mica granitic rocks are also present suggesting a basement for the greenstone that is sialic. Because the greenstone successions are low-grade rocks with tremendous apparent stratigraphic thickness, the original succession is probably in fault relationship against the original basement and this contact has been reintruded by granitic rocks obscuring the original primary tectonic relationships. As such, the granitic rocks provide the best evidence for the composition of the underlying basement.

The fourth basement association is the high-grade gneiss/gabbro/anorthosite association. These rocks have both ortho- and paragneiss. These types of rocks typically have complicated oxygen isotope distributions indicating exchange between the various protoliths over considerable distances (kilometer scales). There are both high and low ^{18}O terranes present in these types of settings.

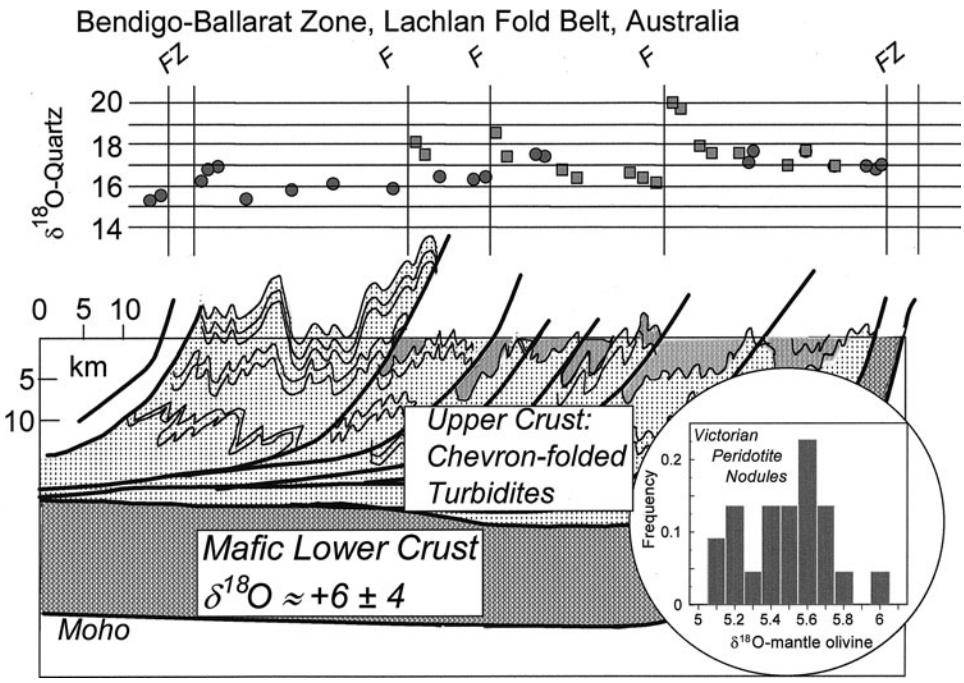


FIGURE 3 A profile through the Lachlan fold belt uses quartz vein $\delta^{18}O$ values as measure of bulk rock oxygen isotopic composition of the crust. Based upon structural analysis the lower crust is probably mafic oceanic crust with an isotopic composition similar to that shown in Fig. 1, but structurally now much more complex (not shown). The Bendigo–Ballarat zone is one of several such zones in the Lachlan fold belt that record a major continental crustal growth event. The new crust is effectively recycled oceanic crust that forms the lower crust. The upper crustal rocks are quite high in $\delta^{18}O$. The inset shows olivine $\delta^{18}O$ values for mantle nodules from the Western Districts Basalts of the Victoria, southeastern Australia. The frequency diagram shows the tenths of per mil type heterogeneities typical for mantle peridotites.

Summing up, the distribution of oxygen isotopes in the continental crust of the Earth indicates that for the most part, low temperature surficial processes impart a high ^{18}O signature on continental crustal rocks so that, on average, the continental crust is enriched in ^{18}O relative to the oceanic crust. Sedimentary rocks deposited from seawater, particularly direct precipitates or biogenic materials, carry the greatest concentration of ^{18}O into the crust. Granitic rocks have higher $\delta^{18}O$ values than primary melts of the mantle and thus have incorporated materials that have previously resided at the surface of the Earth. Granitic rocks therefore are direct chemical probes of lower crustal structure and role of recycling of surface materials to lower crustal and lithospheric levels.

4. Processes Affecting the Oxygen Isotopic Composition of Igneous Rocks

Igneous rocks are mixtures of minerals precipitated from a melt that evolves its $^{18}O/^{16}O$ ratio depending the contrast between the isotopic composition of the precipitating phases and the melt. The $\delta^{18}O$ values of common anhydrous igneous minerals show the following sequence of enrichment: magnetite < olivine < pyroxene < plagioclase < alkali feldspar < quartz. During true closed system igneous fractionation, the ^{18}O-shift of the magma is very small, on the order of tenths of per mil. As a result, the isotopic heterogeneity observed in all lunar igneous rocks sampled to date is on the order of a half of a per mil. Anhydrous igneous fractionation imparts little oxygen isotope heterogeneity into igneous rocks fractionating from basalt to rhyolite. This is commonly observed in some hot spot or ocean island magmatic series.

Even though closed system crystal fractionation does not greatly affect the isotopic composition of igneous systems, the spread in the range of known primary magmatic oxygen isotope composition of igneous rocks is over 10 per mil on Earth. This highlights the role of tectonic processes and crustal recycling and importantly water in the formation of igneous rocks, and in particular more silicic magmas.

In general, the smaller the degree of partial melting in the mantle, the more heterogeneous are the oxygen isotopic compositions of the resulting basaltic melts. This suggests that there are processes that locally affect portions of the lithospheric mantle. In particular, subduction zones represent tectonic regimes where materials carrying surface-derived ^{18}O interact with lithospheric

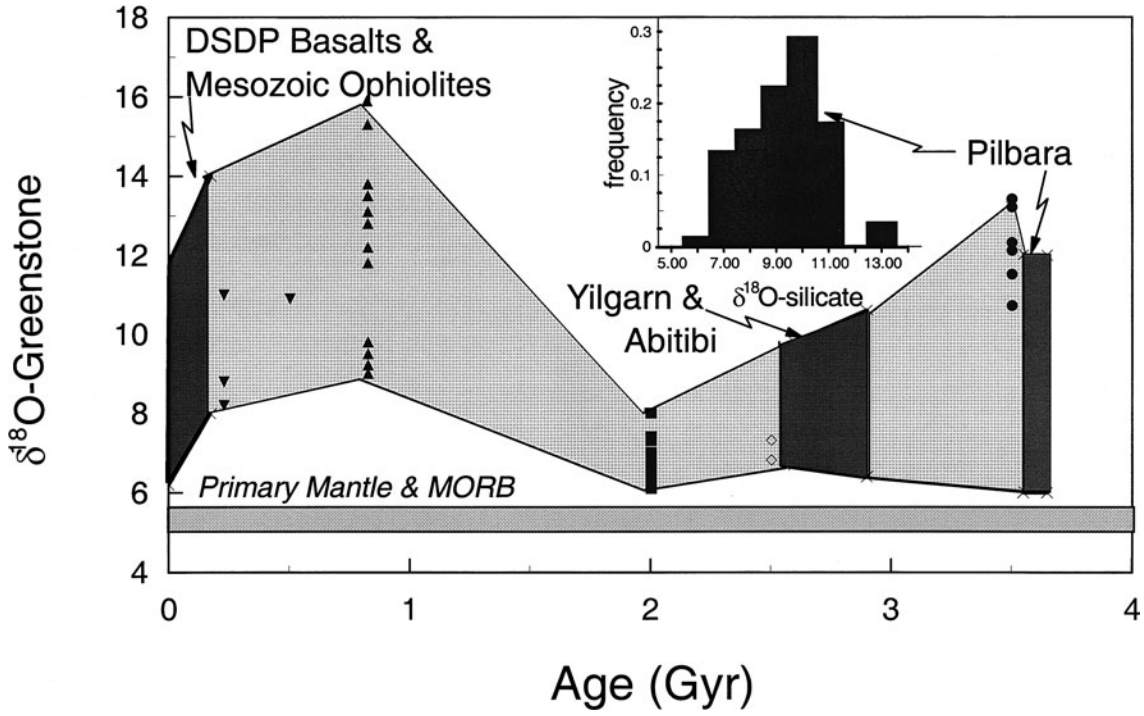


FIGURE 4 Most greenstones and hydrothermally altered pillow lavas are enriched in ^{18}O, independent of their age. The spread in greenstones $\delta^{18}O$ values remains similar throughout Earth history suggesting the hydrothermal fluids and hence the original seawater has remained near its current value over geologic time. Hypothesized swings in the oxygen isotopic composition of seawater to more strongly negative values in the Paleozoic and Proterozoic are not supported by the greenstone data. The dark shaded areas correspond to parts of the rock record where there are many determinations (e.g., the Pibara frequency diagram is based upon over 80 samples).

mantle. Because the most labile material resides in high ^{18}O, hydrated protoliths, the net effect is to enrich small portions of the mantle in ^{18}O. Basaltic rocks such as the alkaline rocks from Italy are examples of regions where the mantle wedge has been enriched in ^{18}O and in ^{87}Sr relative to normal mantle.

There is a several per mil contrast between the average $\delta^{18}O$ value of granitic rocks (+9) and the isotopic composition of melts that could fractionate from partial melts of the mantle to granitic compositions (+6.5 to +7). This suggests that over most of Earth history, granitic rocks incorporate ^{18}O derived from materials affected by near surface fluid–rock interaction and which have been deeply buried by tectonic processes. The highest $\delta^{18}O$ values measured on igneous rocks come from peralkaline or peraluminous granitic rocks indicative of a major contribution from high-^{18}O metasedimentary rocks. Elevated $\delta^{18}O$ values do not always imply a sedimentary source because partial melting hydrothermally altered mafic rocks can result in high ^{18}O ($\approx$+10 per mil) with minor enrichment in ^{87}Sr. Clearly, fluid–rock interaction plays an important role in the evolution of the crust of the Earth and conversely, this interaction affects the isotopic composition of the hydrosphere.

B. Oxygen and Hydrogen Isotopes in the Water Cycle

On the surface of the Earth, the ocean is the dominant reservoir of water consisting of about 97% of all surface water reservoirs that also include (in order of significance) ice caps and glaciers (2%), groundwater (0.68%), meteoric water (0.01%: lakes, rivers, atmosphere, and soils), and the biosphere ($\ll$0.001%). With most of the Earth's surface water trapped in the ocean and other near surface reservoirs, there has been very little loss of water to space and therefore minimal opportunity for the major isotopic fractionation of the oceans by dissociation of water vapor and hydrogen loss through the upper atmosphere.

1. Controls on the Oxygen Isotope Composition of Seawater

A vigorous plate tectonic cycle ensures that the oceans are cycled through midocean ridges on short time scales (10s of millions of years) relative to the age of the Earth (4.6×10^9 years). Sedimentary rocks that carry a surface hydrogen isotope signature are recycled into the mantle at subduction zones. It seems more than a coincidence that

the hydrogen isotopic composition of inferred magmatic waters overlaps with the hydrogen isotopic composition of marine sediments.

In terms of oxygen isotopes, on a planetary scale, the isotopic (and chemical) composition of the ocean is dominated by the competition between two major processes: exchange between mantle-derived reservoirs at midocean ridges and the exchange with silicate rocks at low temperature due to chemical weathering. The change in the isotopic composition of the ocean can be described by a differential equation of the form:

$$\frac{d\delta W_{seawater}}{dt} = \sum_i k_i\left(\delta W_{seawater} - \delta W_{i_{steady\ state}}\right), \qquad (8)$$

where the k_i's represents the exchange rates between the ith reservoir and the ocean and the steady-state δW's represent the delta value the ocean would obtain if the exchange with that particular ith reservoir was the only operative exchange process. This term consists of the isotopic composition of the reservoir minus a bulk fractionation term representing the exchange process between water and rock. The bulk fractionation terms need not be constants and only need to be characterized by mean values over time scales appropriate to the exchange process. The isotopic composition of the oceans tends toward the following composition:

$$\delta W_{seawater} = \frac{\sum_i k_i \delta W_{i_{steady\ state}}}{\sum_i k_i}. \qquad (9)$$

Perturbations away from the steady-state isotopic composition decay away as $\exp -(\sum k_i t)$; e.g., Fig. 5. An analysis of the rates of exchange indicates that the seafloor exchange process should dominate the oxygen isotopic composition of seawater over geologic time; i.e., the oxygen isotopic composition of seawater is buffered on 10s to 100s of millions of years time scales. In other words, the time scale for this global cycle is much shorter than the age of the Earth. The quasi-steady-state composition is the weighted mean of the target values for each part of the cycle.

If the rate constant is 0.0125/Ma, the residence time for the midocean ridge black smoker cycle is 80 Ma. Similarly, if the rate constant is 0.0025/Ma, the continental weathering input residence time is 400 Ma. These rate constants are calculated from the volumetric spreading and weathering rates normalized to moles of oxygen exchanged per mole of seawater oxygen per unit time. If the midocean ridge process is driving seawater toward the black smoker vent fluid (e.g., +1 per mil) and chemical weathering is driving the oceans toward −12, the steady state is calculated from the k terms in Eq. (9). The mixing proportions, f, are

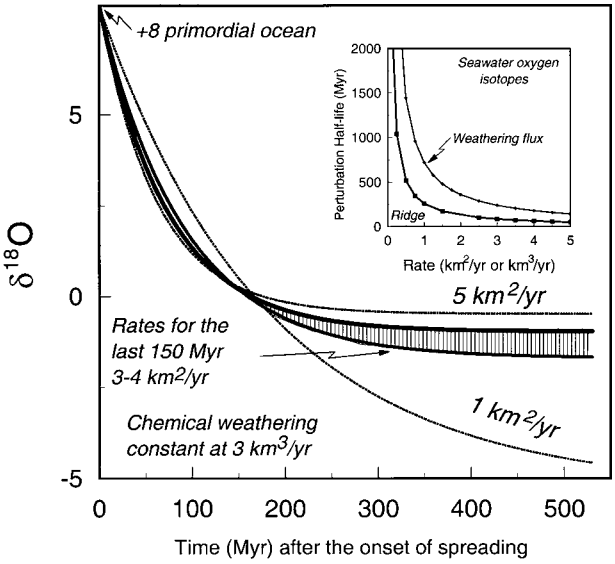


FIGURE 5 The time constants for the obtaining quasi-steady-state for the oceans. The calculation assumed a +8 initial ocean whose isotopic composition is driven by the competition between chemical weathering and hydrothermal activity on the seafloor. The inset shows the half-life of any perturbation in the oxygen isotope composition of the oceans. For geologically plausible choice of rates, the ridge process dominates the oxygen isotope composition of the oceans.

$$f_{mor} = 0.0125/(0.0125 + 0.0025) = 0.83$$
$$f_{cw} = 0.0025/(0.0125 + 0.0025) = 0.17.$$

Note the values for the f's sum up to 1. Then target steady-state value is $\delta^{18}O = 0.833 * 1 + 0.167 * (-12) = -1.2$, where −1.2 per mil is the weighted average of the mixture maintained by the dynamical process driven by plate tectonics. The value is not far from the value seawater would achieve if the ice caps were added back into the ocean.

This result for oxygen is in contrast to that for Sr isotopes. For the Sr system, the Sr isotopic composition of the oceans is more variable over geologic time and the time constants for exchange are much shorter (million-year scale). However, normalizing all of the rates to global spreading rate (km^2/year or km^3/yr of new crust or global volumetric weathering rates, km^3/yr of continental rocks dissolved) shows that the same tectonic rates account for both isotopic systems. The difference in the time scales results from the concentration differences between Sr and O in rocks (100s ppm and 45 wt%, respectively) and in fluids (8 ppm in seawater, 89 wt%, respectively).

Tectonic processes ultimately drive long-term geochemical cycles so that molar elemental fluxes for true long-term global cycles run at rates of cubic kilometers per year of material processed, i.e., typical tectonic rates.

Depending upon the concentration of the element of interest in the ocean, the relevant time scales can vary by orders of magnitude. Short-term geochemical cycles may operate at time scales that are insignificant when compared to the time scales involved with the longer-term geologic cycles. These fast cycles represent perturbations that operate around the means of steady-state values of the longer cycles. The per mil level changes in the isotopic composition of seawater driven by glacial cycles is a classic example of this difference.

Stable isotope ratios determined on ice core samples and tests of marine plankton indicate that ice sheets build up gradually and retreat rapidly. Glacial reservoirs of ice involve as much as 10^4 to 10^5 km^3 of water fractionated per year (see the following for the mechanism) considering that 10^8-km^2 areas are affected by continental glaciation and typical global precipitation rates are on the order of 1 m/year. Because the waxing and waning of ice sheets appear to occur on Milankovitch time scales ($\approx 10^4$ yr), rapid 1 per mil excursions in the isotopic composition of the global ocean occur with little or no impact on the long-term geologic cycle. The glacial perturbations are reversible and therefore average out to zero before enough time has elapsed for the long-term cycle to respond to any particular glacial advance or retreat.

2. Meteoric Water Cycle

The constancy of the oxygen isotopic composition of the global ocean over geologic time scales has important consequences for the meteoric water cycle and paleoclimatology. Virtually all of the water in the atmosphere involved in weather systems ultimately has an oceanic source. As air masses move away from the ocean, they cool along atmospheric adiabats. Ultimately, condensation of water vapor occurs and this is related to the temperature structure of the atmosphere. If the water is perfectly removed from the system, the isotopic composition of the remaining air mass is depleted in both hydrogen and oxygen isotopes by Rayleigh distillation.

After large amounts of the original water have been removed from the atmosphere the isotopic composition of the residual vapor and hence the rain precipitated produce some of the most depleted isotopic compositions for hydrogen and oxygen isotopes (5% for ^{18}O and 40% for deuterium). Even though the fractionation factors for hydrogen and oxygen have different temperature dependencies, the fraction of vapor remaining is the same for both during the condensation process. As a result, the isotopic composition of the remaining vapor and the precipitation can be mathematically related by solving for the fraction of vapor or liquid in the system. The resulting relationship is remarkably linear:

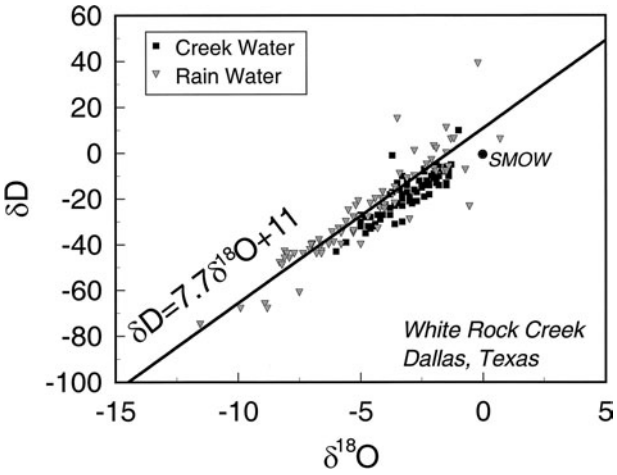


FIGURE 6 Meteoric waters and creek runoff for White Rock Creek, a small drainage in Dallas County, TX. Note that the creek waters also lie close to meteoric water line and exhibit less variation. Shallow groundwater flow into the creek delivers a "smoothed" weighted average precipitation composition slightly modified by evaporation. Each local area has its own meteoric water line.

$$\delta \mathrm{D} \approx 8\delta^{18}\mathrm{O} + 10. \qquad (10)$$

Natural meteoric waters (e.g., see waters for Dallas, TX, Fig. 6) confirm the predicted relationship spanning a range of more than 50 per mil in oxygen and 400 per mil in deuterium. The most isotopically depleted samples are found in the polar regions and the most isotopically enriched samples are found in arid regions, particularly from bodies of water subject to extreme evaporation.

As an air mass rises and cools, the saturation vapor pressure of the atmosphere decreases with decreasing temperature. Eventually, the air mass reaches the vapor saturation curve and liquid water begins to form. As the air mass cools further, the vapor pressure of water continues to drop and more liquid or solid water condenses and ultimately falls back to the surface as precipitation. This process imparts a temperature dependence on the isotopic composition of precipitation that is particularly important for precipitation related to frontal systems that sweep across the continents.

Many years of International Atomic Energy Agency network data show that the average isotopic compositions of meteoric water correlates with mean surface temperatures. This works particularly well when the mean surface temperature drops below 15°C. Above 15°C, something called the "amount" effect becomes dominant and the relationship between surface temperature and the isotopic composition of rain breaks down (Fig. 7). This latter scenario is more common in the tropics and summer storm activity such as monsoons.

For mean annual or mean monthly temperatures below 15°C, the correlation between mean surface temperature

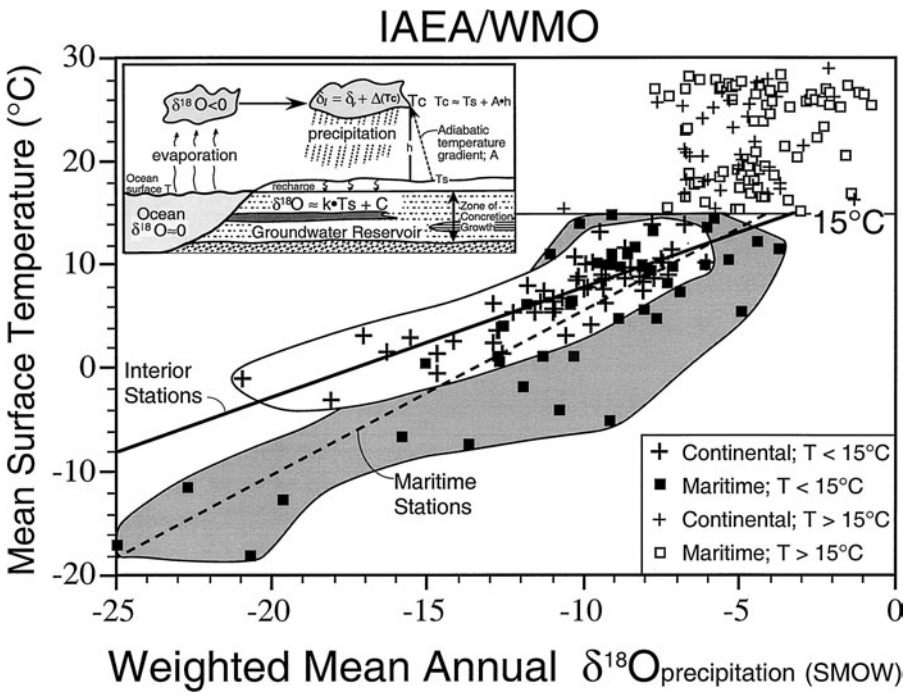


FIGURE 7 The weighted mean annual $\delta^{18}O$ values of precipitation compared with mean surface temperature. The correlation is best observed for stations where the mean annual temperature is less than 15°C. The inset shows a schematic representation illustrating why there should be a correlation between surface temperature and precipitation when the primary fractionation step is occurring during liquid water nucleation in the clouds. This type of correlation (slightly less that 1°/1 per mil change in mean meteoric water) is the basis for proxy paleotemperature estimates; e.g., ice cores).

and the mean isotopic composition of the precipitation is striking. Because general mean air temperature decreases from the equator to poles on a global scale, the so-called latitude effect on the isotopic composition of meteoric waters is a consequence of the temperature structure of the atmosphere and the mechanics of cloud formation and precipitation. The decrease in isotopic composition of meteoric water as a function of elevation also reflects the lapse rate of temperature in the atmosphere on a more local scale.

3. Subsurface Fluids

Subsurface fluids are (1) shallow groundwaters similar in isotopic composition to meteoric waters, (2) formation waters with a complicated provenance, (3) geothermal fluids, (4) metamorphic waters, (5) magmatic water, or (6) juvenile water still outgassing from the interior of the Earth. Regarding the latter, it would be nearly impossible to identify juvenile water unless its hydrogen isotope composition is dramatically different from water dissolved in magmas erupting today. There has been enough cycling of hydrogen into the lithosphere at subduction zones to suggest that the upper mantle has been contaminated with surface derived hydrogen.

Shallow groundwaters typically have an isotopic composition very similar to the local average meteoric water (e.g., Fig. 7, White Rock Creek is a gaining stream) and are only modified chemically through reactions that occur during low temperature weathering. The major cation and anion chemistry of these groundwaters depends upon the reactivity of the host rocks. For example, in areas where active serpentinization is occurring, shallow groundwaters exhibit very high pH (>10) while retaining the local meteoric water stable isotopic composition.

Clay minerals in soils typically have hydrogen and isotopic compositions that parallel the meteoric water line, so much so that in a graph of δD against $\delta^{18}O$, the array mapped out by such determination is called the clay line (e.g., the kaolinite line of Fig. 8). The clay lines are subparallel on a global scale to the meteoric water line. In regions where the minerals are growing in a system open to fluid flow and connected to the surface, the minerals form in isotopic equilibrium with the surface fluid; hence their δD and $\delta^{18}O$ values parallel the meteoric water line. These systems behave as if the mole fraction of water approaches 1.

In contrast, formation waters can be tapped from zones where the effective mole fraction of water is very small so

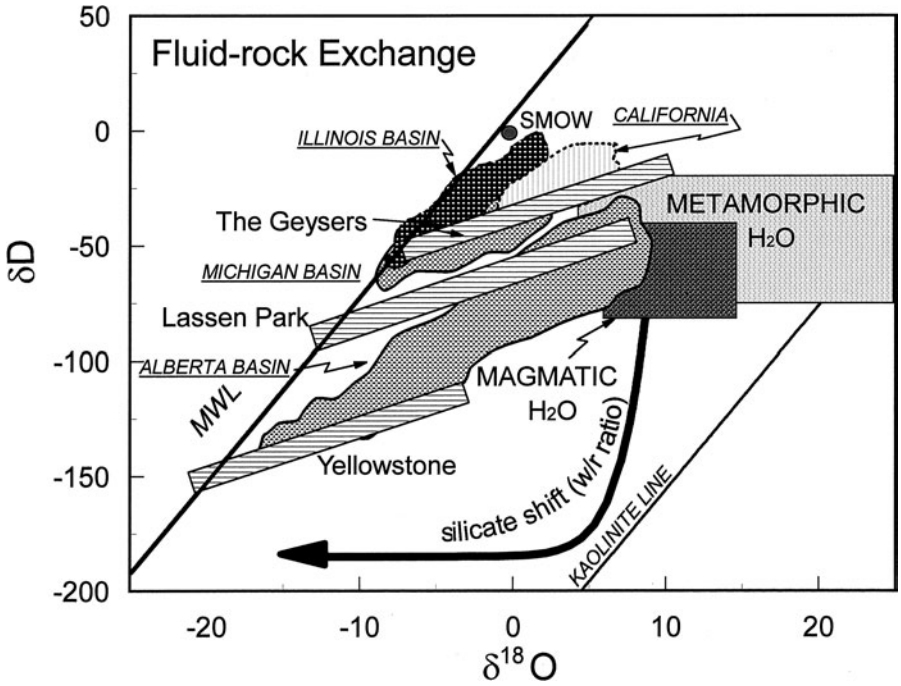


FIGURE 8 The fields for a sampling of different crustal fluids compares measured basinal fluids (underlined labels) with fluids from three geothermal areas (Yellowstone, Lassen Park, and the Geysers, an electric power generating field in California). Schematic fields for metamorphic fluids and primary igneous rocks are calculated from rock compositions. Nearly all of these fluid–rock interaction systems show evidence for the circulation of surface fluids to great depth. The "^{18}O shift" refers to the locus of points starting at the meteoric water line plotting toward the right in the diagram with either increasing temperature or a greater approach to isotopic equilibrium in the rocks. Also shown is a typical trajectory a granitic rock might follow during subsolidus exchange with a circulating meteoric fluid. Such rocks have been observed extensively in rifts zones and a great granitic belts such as the Cordilleran batholiths. The kaolinite line is shown to illustrate how clay minerals formed during weathering parallel the meteoric water line.

that isotopic and chemical exchange become more important in determining the signature of the fluids. In terms of simple mass balance, the water to rock ratio or the related parameter the integrated fluid flux is important for these types of fluids. The former parameter is dimensionless, whereas the latter has units of length (cm^3 of water per cm^2). The characteristic water–rock ratio times the length scale of the flow system gives the integrated volumetric fluid flux.

Formation waters are mixtures of original trapped fluids such as seawater and fluids modified by rock–water interaction (hydration or dehydration reactions or crustal transport). The latter fluids are the most difficult to characterize because the isotopic signature and the major element chemistry in the dissolved load depend upon the fluid migration path and the pressure and temperature conditions encountered along the path. For example, oil field brines show a deuterium isotope dependence that appears to be correlated with latitude indicating the meteoric waters have recharged some of these subsurface basins (Fig. 8). In other parts of these oil field basins, major element data suggest that fluids derived from the original trapped seawater are present, now strongly ^{18}O-shifted and in some cases even D-shifted from their original compositions.

Geothermal fluids typically lie on straight line data point arrays between a rock buffered oxygen isotope composition and the local meteoric fluid (Fig. 8). Because hydrogen is far less abundant in rocks, the fluids are strongly ^{18}O-shifted but shifted little from their surface δD value. Conversely, hydrothermally altered subvolcanic rocks often exhibit a hydrogen isotope shift toward D-depleted meteoric compositions before their oxygen isotope ratios show any major effects (see the curve on Fig. 8, the arrow shows the direction of silicate isotopic evolution along the line). Larger haloes of D depletion overlie smaller haloes of ^{18}O depletion. Zones of maximum depletion represent regions of high integrated fluid flux. At the surface, the subsurface hydrothermal alteration manifests itself as surface hot springs with variable amounts of isotopic shift away from the local groundwater. All of the stable isotopic data suggest that meteoric fluids penetrate several kilometers into the crust of the Earth and affect large areas of rock centered around plutonic heat engines.

Metamorphic fluids are usually presumed to be in isotopic equilibrium with the local host rocks at pressure and temperature conditions at the time of metamorphism. Through fluid inclusions it is possible to analyze the hydrogen isotopic composition of trapped metamorphic fluids. The δD values from these fluid inclusions exhibit ranges comparable to those expected from fluids equilibrated with the hydrous minerals in the rocks. In fact, the hydrogen isotope ratios of marine sediments, metamorphic rocks, and unaltered igneous rocks are remarkably similar suggesting that the water sampled by these rocks ultimately is surface fluid bound in hydrous phases tectonically cycled into the source regions for metamorphic and igneous rocks.

C. Oxidation/Reduction Reactions: Hydrogen, Carbon, and Sulfur Isotopes

Hydrogen, carbon, and sulfur are elements that participate in oxidation/reduction reactions. In contrast to oxygen, which makes up over 40 wt% of the crust, these elements are present in trace quantities on a lithospheric scale. For example, hydrogen and methane are important components in oxidation/reduction reactions. There are large fractionations associated with the reduction of water in silicate mineral structures to produce hydrogen gas. Hydrogen–water or hydrogen–methane fractionations remain large (100s per mil) even in metamorphic and igneous regimes.

Carbon and sulfur participate in oxidation/reduction reactions where the valence of the element can change from positive to negative (e.g., CO_2 going to CH_4 or C^{4+} in carbonate versus C^{4-} in organic matter). These types of reactions involve large fractionations of the stable isotope ratios so that the natural variability for sulfur and carbon isotopes approaches 100 per mil. Exchange of sulfur isotopes between sulfur bearing compounds can be quite sluggish under crustal conditions so that isotopic nonequilibrium is common.

The energy released from oxidation/reduction reactions provides energy for life. At the surface of Earth and within the crust of the Earth, bacteria utilize these reactions and form the bottom of any food chain within an ecosystem. The names of the bacteria (e.g., *Methanococcus*) provide some insight into the reactions that provide the source of energy; these include reactions involving organisms and chemical feedstock such as carbon dioxide and water, sulfide, sulfate, organic matter, and methane. Biogenic oxidation reduction reactions usually leave behind large fractionations in carbon and sulfur isotope systems (Table I).

Because of the presence of free oxygen in the atmosphere, the hydrosphere and shallow lithosphere are more oxidizing than the conditions that would exist in the absence of life. The source regions of most magmas are more oxidizing than the conditions necessary to separate iron metal from silicate melts. Because the iron-rich core did separate from the Earth's mantle, this requires that (1) there is a pressure dependence on the oxygen fugacity necessary to separate iron from silicate, (2) a late veneer of more oxidizing material was added to the Earth as a result of late bombardment of the Hadean Earth, or (3) that recycling of volatiles into the lithosphere has been an important process during Earth history. Stable isotopes provide some insight on the latter possibility.

1. Hydrogen in the Lithosphere

The ocean represents the major reservoir of hydrogen in lithosphere. Because of the presence of ice caps, the ice-free δD of the ocean is about -10 per mil. Water outgassing from primary magmas is not –10 per mil, but rather in the range -40–-80 per mil; this range coincides with the range for marine sedimentary rocks or metamorphic rocks. Some hydrous minerals from mantle peridotites suggest a reservoir of water between $-23 < \delta D < -45$ suggesting a more enriched source of D in the upper mantle. There may be a correlation between enriched δD values and Fe^{3+} in hydrous phases from these mantle regions suggesting that the reduction of water liberates ^{2}H-depleted hydrogen and leaves behind D-enriched hydrous minerals. These data strongly suggest that lithospheric hydrogen sampled by igneous rocks is primarily recycled surface waters. This conclusion is certainly plausible considering the early outgassing history of the Earth due to planetary scale heating that occurred as a result of accretion, core formation, and/or early catastrophic impact history.

2. Carbon in the Lithosphere

The atmosphere of the Earth has a δ^{13}C value approximately -5 to -7 per mil that is now probably perturbed by the addition of anthropogenic carbon dioxide derived from the combustion of fossil fuel. Crustal carbon (3 parts carbonate and 1 part organic or reduced carbon) is also approximately -5 per mil. Primary magmas also seem to outgas on average -5 per mil carbon, a value similar to that for carbonatite and the peak of the diamond δ^{13}C distribution. However, diamonds exhibit a wide range of δ^{13}C values, $-35 < \delta^{13}C < +2$. Although oxidation or reduction of fluid species CH_4 or CO_2 associated with silicate magmas could conceivably produce the observed variation by Rayleigh separation (distillation), in light of plate tectonics, it is more plausible that the δ^{13}C values of diamonds reflect recycling of crustal carbon. The association of diamonds with kimberlites and lamproites, both of which contain mantle inclusions thought to be recycled crustal

material, also supports this hypothesis. In addition, the carriers of diamonds to the surface, the host rocks themselves, exhibit elevated $\delta^{18}O$ values and $^{87}Sr/^{86}Sr$ values (e.g., lamproites), and this also supports the inference that carbon in the lithospheric mantle is largely recycled crustal carbon.

3. Sulfur in the Lithosphere

Meteoritic sulfur has very uniform $\delta^{34}S$ values ($\approx$0 per mil). Assuming that the mantle of the Earth inherited this primordial sulfur isotope composition, sulfur in the mantle (stable as sulfide because of the oxidation state of the mantle) should be relatively uniform in $\delta^{34}S$. Melting of the mantle should result in a net transfer of sulfur to the crust of the Earth where under surface conditions it partitions into oxidized and reduced species. Recycling of crustal sulfur could transfer crustal type isotopic heterogeneities back into the mantle. Limited measurements on uncontaminated primary mantle materials suggests that the mantle may indeed have meteoritic $\delta^{34}S$ values. Crustal sulfur estimates range from approximately $+2$ to $+9$ per mil, clearly different from presumed primordial $\delta^{34}S$ values. More recent bulk crustal estimates tend toward presumed mantle values. Crustal sulfur is a mixture of reduced sulfur and seawater-derived sulfate with a modern $\delta^{34}S$ value approximately $+20$ per mil. Granitic rocks presumed to be probes of the their crustal source regions cluster around $+7$ per mil for samples with the highest concentrations of S (>100 ppm S) in closer agreement with higher values for crustal $\delta^{34}S$.

Mantle-derived igneous rocks such as gabbros exhibit $\delta^{34}S$ values that require assimilation of heterogeneous crustal materials or heterogeneities in the isotopic composition of mantle-derived sulfur. Several layered intrusions (e.g., the Duluth Complex, the Muskox, the Noril'sk, and the Sudbury Complex) exhibit positive $\delta^{34}S$ values ($0 < \delta^{34}S < +17$) suggestive of assimilation of crustal sulfate. In contrast, the Bushveld exhibits negative $\delta^{34}S$ values (-6 to -9 per mil). If the Earth has a $\delta^{34}S$ value similar to meteoritic sulfur and the crust is somewhat enriched in ^{34}S, where is the complementary negative $\delta^{34}S$ reservoir? Presumably, this reservoir resides in the lower crust or some process preferentially recycles ^{34}S-depleted biogenically reduced sulfur into the mantle.

III. GLOBAL CYCLES AND SECULAR CHANGE

A. Long-Term Coupled Carbon–Sulfur Isotope Variations

For at least half of Earth history, the Earth has had an oxygen-rich atmosphere. The onset of this condition, unique in the solar system, is still controversial. The remarkable balance between reduced/organic carbon and carbonate reservoirs over Earth history is suggestive of early evolution of an oxygen-rich atmosphere. A comparison of hydrothermal carbonate $\delta^{13}C$ values from 3.5-Gyr-old Pilbara pillow lavas with carbonate alteration from Deep Sea Drilling Project basalts illustrates the coincidence of values (Fig. 9). Because $\Delta^{13}C$ HCO_3^-–$CaCO_3$ fractionations are relatively insensitive to temperature, hydrothermal carbonate is a good proxy for sedimentary $\delta^{13}C$ (marine carbonate rocks are rare in the Archean). Organic matter combusted from cherts in the Pilbara has an average $\delta^{13}C \approx -35$ typical for rocks of this age and metamorphic grade. The sulfur isotope record is less complete

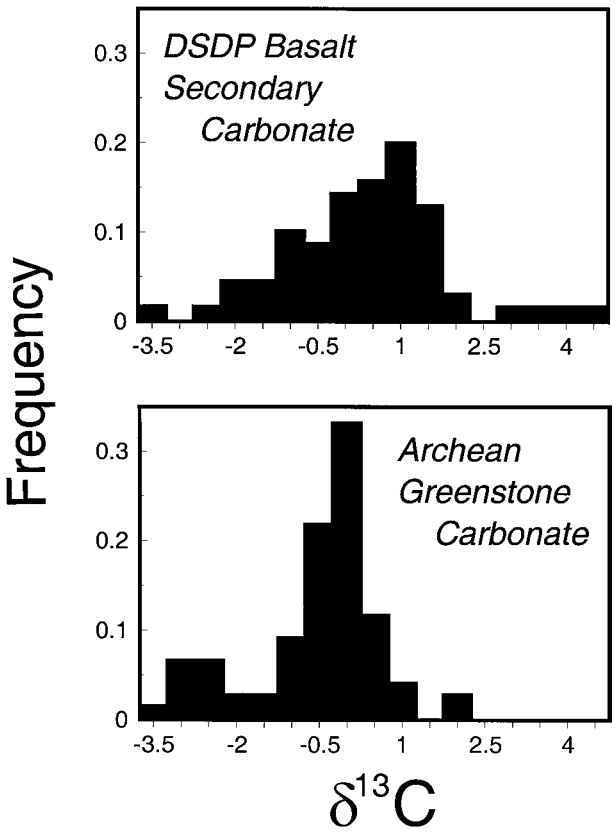


FIGURE 9 Hydrothermal carbonates from Archean greenstones of the Pilbara are compared against secondary carbonates in basalts sampled by the Deep Sea Drilling Project. Measurements of ancient organic matter demonstrate that the organic carbon reservoir has remained within a restricted range throughout Earth history with mean values of about -26 for much of Earth history and slightly more ^{13}C depleted values (-30 to -35) for some Archean rocks. The high-$\delta^{13}C$ values of the hydrothermal carbonate suggests that at the time of seafloor metamorphism (>3.4 Gyr ago), the balance between the carbonate and reduced carbon reservoirs had been obtained for the long-term carbon cycle. Had this balance not been obtained the peak of the Archean distribution should be closer to -5 to -7 per mil.

because evaporites particularly gypsum and anhydrite are not well preserved in the rock record. The oldest sulfates are commonly barite which is less soluble under crustal conditions.

An important result of combined sulfur and carbon studies on Phanerozoic rocks on both isotopic ratios and elemental ratios is the antithetical variation between carbon and sulfur in the rock record. Abundant organic carbon burial seems to coincide with times of sulfate deposition. Abundant carbonate deposition accompanies times of reduced sulfur burial or pyrite deposition. High-$\delta^{34}S$ values coincide with times of lower-$\delta^{13}C$ values of marine carbonates (Fig. 10). Because it is not common to measure both isotopes in the same section or even have the coexisting minerals in the same specimens, these correlations are based upon long-term averages and synthesis of many sedimentary sections for a given time interval. Nevertheless, these observations suggest that whatever mass balance equation might be employed to describe the combined carbon/sulfur cycle, that equation contains oxidized and reduced species on *both* sides of the material balance relationship.

The inverse relationship for isotope ratios and concentration ratios allows the material balance relationship to be written on a constant oxygen basis, i.e., an increase in sulfate burial should be accompanied by increase in organic carbon burial. Under the constant oxygen assumption four reactions can be combined involving the following processes: (1) the precipitation or dissolution of calcium sulfate, (2) photosynthesis or respiration, (3) oxidation of pyrite or the reduction of sulfate, and (4) carbonate–silicate reaction (weathering/metamorphism). The grand expression is

$$4FeS_2 + 7CaMg(CO_3)_2 + CaCO_3 + 7SiO_2 + 15H_2O \Leftrightarrow 8CaSO_4 + 2Fe_2O_3 + 7MgSiO_3 + 15CH_2O. \quad (11)$$

Reaction (11) can be written with atmospheric carbon dioxide explicitly in the reaction:

$$CO_{2(g)} + 2Fe_2O_3 + CaSO_4 + MgSiO_3 + 15CH_2O \rightarrow 4FeS_2 + 8CaMg(CO_3)_2 + 8SiO_2 + 15H_2O. \quad (12)$$

Reaction (12) written in this direction describes conditions that might obtain during rapid spreading and plate convergence, high sea level, high partial pressures of atmospheric CO_2, and enhanced volcanic and metamorphic return of carbon dioxide to the atmosphere. With the continents flooded, there is a net return of sulfate to the ocean and deposition of pyrite with elevated $\delta^{34}S$ values. There is a net return of ^{13}C-depleted material to the oceans to be

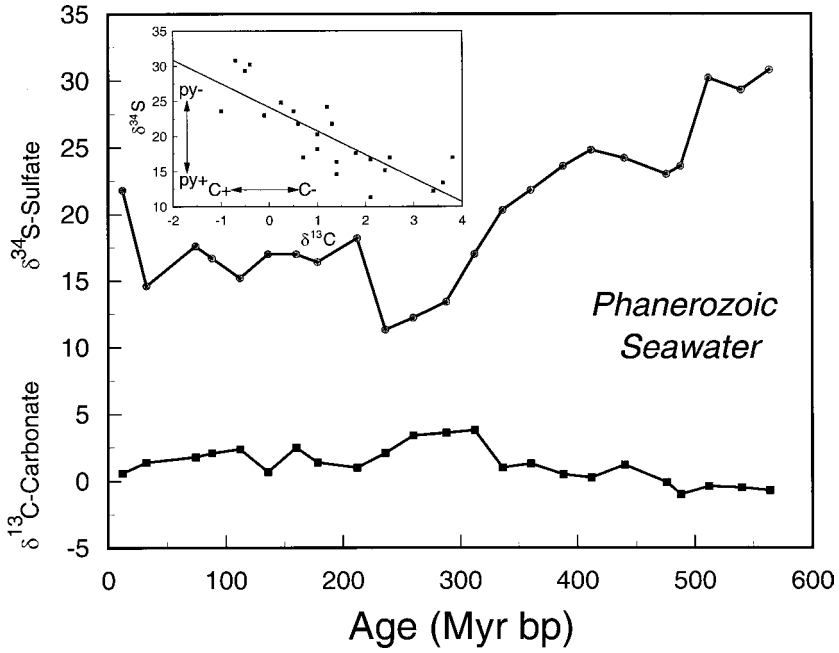


FIGURE 10 The Phanerozoic change curve for carbonate and sulfate minerals precipitated from seawater or seawater-derived fluids. A major secular decline in the $\delta^{34}S$ of marine sulfate and enrichment in the $\delta^{13}C$ in carbonates climaxes with a major glaciation in Carboniferous/Permian. The isotopic ratios are clearly coupled from the secular change curves; this is more dramatically displayed in the inset that shows the variation of $\delta^{34}S$ against $\delta^{13}C$. The negative slope suggests that pyrite (reduced sulfur) and organic carbon (reduced carbon) are on the opposite sides of any material balance for the oceans. Burial of ^{13}C-depleted organic carbon (C−) drives the remaining carbonate $\delta^{13}C$ more positive while oxidation of ^{34}S-depleted pyrite (py+) drives the $\delta^{34}S$ more negative.

precipitated as carbonate with lower-$\delta^{13}C$ values. The Cretaceous greenhouse time would be an example of this scenario.

At times of slow spreading, reaction (12) goes to left. Written going to the left, the equation describes low stands of sea level, small midocean ridge volume, lower rates of convergence and arc magmatism (and hence reduced CO_2 outgassing), and low partial pressure of atmospheric carbon dioxide. Pyrite dissolution produces acid that dissolves carbonate and produces sulfate to be precipitated as anhydrite or gypsum. Organic carbon is buried in the ocean and the $\delta^{13}C$ values of the remaining precipitated C as carbonate swing positive.

These tendencies in the secular change record of sulfur and carbon suggest that there is a strong tectonic control on the variation of organic C/S ratios and the δ values of the two elements. The link between global spreading rates and sea level is a critical boundary condition in determining habitats for living things, atmospheric CO_2 concentration, weathering rates, depositional environments, and in conjunction with the Sun a major factor influencing climate.

B. Oxygen Isotopes and Paleoclimatology

The isotopic composition of material precipitated from seawater yields a direct estimate of the paleotemperature provided that the sample does not exchange subsequent to its formation and that the isotopic composition of the fluid is known. The latter is particularly difficult for ancient samples. The uncertainty and the local variation of the isotopic composition of seawater (at the per mil level) guarantee an uncertainty of at least $\pm 3^{\circ}C$ in typical paleotemperature estimates. The small shift in the oxygen isotopic composition of the ocean that occurs between glacial–interglacial cycles obscures the temperature signal so that the measured numbers provide a combined ice volume/temperature signal. Nevertheless, the periodicity observed in the marine cores is the best record of glacial advance and retreat.

Using the isotopic composition of meteoric waters to infer climatic conditions is a less direct means to get at paleotemperatures. This is particularly so when the composition of the fluid represents either an average of many events (ice cores) or an average value of some fluid in equilibrium with a solid phase such as a soil carbonate, a concretion, or tooth enamel from a terrestrial herbivore. Because of the uncertainties in the correlation between surface temperature and isotopic composition of precipitation and in the formation conditions of the proxy phase (e.g., carbonate), indirect estimates of paleotemperatures have an uncertainty somewhat greater ($\pm 5^{\circ}C$).

1. The Marine Record: Tertiary and Quaternary Oxygen Isotopes

A comparison of the record of marine and benthonic foraminifera provides some of the best evidence for the onset of icehouse conditions in the Tertiary (Fig. 11). In today's oceans latitudinal temperature gradients from the poles to the equator coupled with a longitudinal distribution of continents and the isolation of the Antarctica set up conditions for strong thermohaline circulation, i.e., dense, cold seawater sinks in polar regions and flows toward the equator.

The difference in the oxygen isotope composition between benthonic forams (bottom dwelling organisms) and planktonic forams (surface water organisms) is a measure of the strength of the thermocline or the contrast between bottom water temperatures and surface water temperatures. Starting in the Eocene, high latitude surface water and bottom water temperatures begin to cool, perhaps the result of the isolation of the Antarctic continent as the great south polar continent of the Mesozoic broke up with the northward migration of Australia, India, and Africa away from Antarctica. Dramatic changes occur at the end of the Eocene, toward the end of the early Miocene, and again in the Pliocene.

The Quaternary record from marine foraminifera exhibits periodicities consistent with some type of orbital and precessional forcing. Particularly over the last 600 k.y., the frequency content and the amplitudes of the oxygen isotope record are particular striking, showing forcing with 23, 41, and 100-k.y. periodicity (Fig. 11). The typical

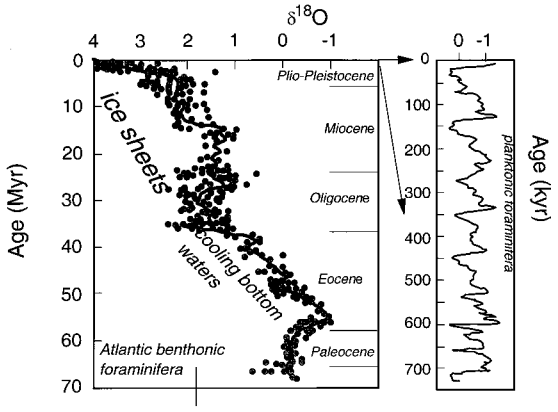


FIGURE 11 The Cenozoic secular change record for the $^{18}O/^{16}O$ ratios of carbonates shows the change in bottom water temperature recorded by benthonic foraminifera in the Tertiary. As Australia mores northward, Antarctica becomes more isolated at the south pole. As the Atlantic grows to become a major ocean, the longitudinal distribution of continents coupled with the isolation of Antarctica results in a reorganization of global ocean circulation. The Pleistocene planktonic foram record shows the signal induced by Milankovitch cycles in a combined ice volume-temperature signal for the last 700,000 years.

cycle begins with a sharp swing toward lower-δ^{18}O values followed by a more gradual increase in δ^{18}O punctuated by shorter period cycles culminating in a maximum δ^{18}O value at the end of the 100-k.y. cycle followed by a rapid excursion to lower-δ^{18}O values. δD and δ^{18}O determinations on ice cores coupled with measurements of greenhouse gas concentrations in trapped gas inclusions show that the δ^{18}O variation in foraminifera are tracking similar global changes related to large scale climate change.

2. The Vostok Ice Core Record

The correlation between surface temperature and the isotopic composition of snow forms the basis of ice core paleothermometry, a spectacular example of which is the record of the Vostok ice core (Fig. 12). This record has now been extended out to approximately 450,000 years. It is a record of oxygen and deuterium isotope concentrations in ice originally precipitated as snow and a record of methane and carbon dioxide concentrations in trapped gases as well as other trace elements recorded by the contamination of ice with atmospheric dust particles. The oxygen and deuterium isotope results corroborate the variations observed in the oxygen isotope record of deep sea cores and record the oscillations between glacial and interglacial periods of the last part of the Quaternary. The trapped CO_2 concentrations illustrate background variations in atmospheric CO_2 before the perturbations induced by anthropogenic emissions related to fossil fuel combustion and deforestation.

C. Carbon Dioxide in the Atmosphere

1. Geologic Variation

Inventories of crustal reduced carbon and sulfur in sedimentary rocks resulted in a series of geochemical models that try to predict distributions of sedimentary rocks. The secular change in their chemical and isotopic compositions should be a function of some intensive variable of the Earth system (e.g., paleotemperature, pressure of oxygen or carbon dioxide in the atmosphere).

Carbon dioxide is of particular interest because it is a greenhouse gas and in combination with water makes carbonic acid which is important in weathering reactions. On long-term geologic time scales, carbon dioxide is removed from the atmosphere through weathering reactions that ultimately convert silicate rocks into carbonate rocks. These carbonate rocks are recycled at plate boundaries and decarbonation reactions return carbon dioxide back to the

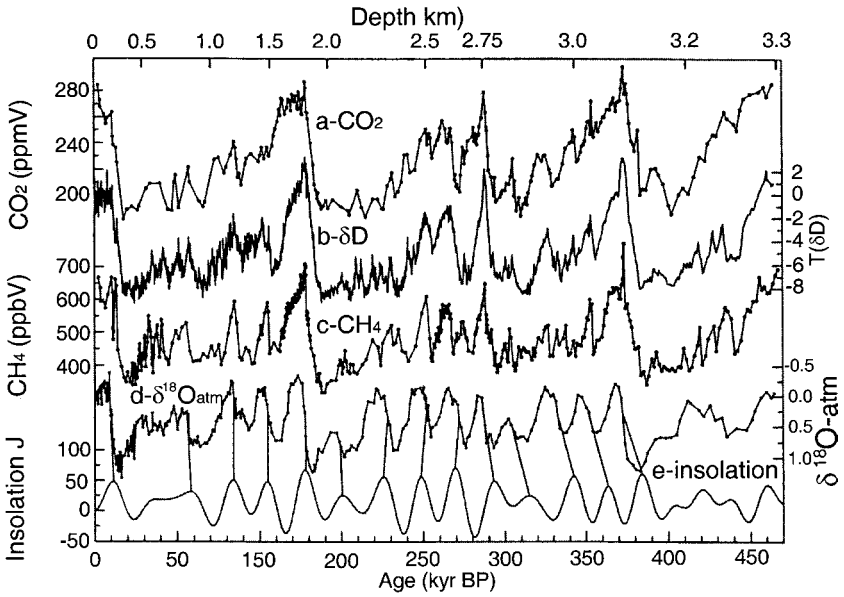


FIGURE 12 The Vostok ice core record now records 450,000 years of climate change. Curve a shows the concentration of carbon dioxide gas in parts per million by volume for trapped atmospheric gas bubbles and defines the preindustrial ranges for this gas; the average is nearly 100 ppmV below current levels. Curve b shows the deuterium isotope record of the ice transformed into a relative temperature scale showing the change in mean temperature between glacial and interglacial periods. Parts of five 100,000 cycles are preserved, three complete cycles and two partial cycles. Curve c shows the preindustrial ranges for another greenhouse gas methane, also below current levels. Curve d shows the δ^{18}O value of the trapped atmospheric oxygen. This measurement is used to infer the magnitude of the δ^{18}O shift of the ocean due to ice volume changes. Curve e shows the insolation at high latitude, a measure of the solar flux change as a result of the precession of the Earth's spin axis and the slight eccentricity of its orbit.

atmosphere. These processes are often written in the short hand form called the carbonate silicate cycle:

$$CO_2 + CaSiO_3 \Leftrightarrow CaCO_3 + SiO_2. \qquad (13)$$

Reaction (13) going to the right expresses the near surface cycle of chemical weathering followed by deposition of carbonate and silica in the oceans. Reversing the reaction (going to left) expresses metamorphism in the crust with carbon dioxide returning to the surface dissolved in a metamorphic fluid or in a magma. The existence of liquid water on the surface and in the crust of the Earth along with the retrograde solubility of carbonate ensures that carbon dioxide will not be the dominant constituent of the atmosphere. As a result of these processes, the Earth has had a nitrogen-rich atmosphere for most of its history.

Remarkably, forward models of the global cycles based upon the distribution of surficial rocks predicted major changes in the partial pressure of carbon dioxide throughout the Phanerozoic. In particular, concentrations as high as 15 times present-day atmospheric values were inferred for parts of the Paleozoic. Carbon isotope records represent the best proxy record of carbon dioxide partial pressures in the atmosphere. The peaks in the bimodal distribution of inferred atmospheric carbon dioxide concentrations corresponded to periods of mostly global greenhouse conditions. The Carboniferous to Permian glaciation along with the current glacial epoch (global icehouse or cool zones) correspond to periods of inferred carbon dioxide concentration minimums.

Figure 13 shows the predicted concentrations of atmospheric CO_2 with data points determined using the solubility of carbonate in the iron hydroxide mineral goethite. When goethite forms in the soil weathering zone, it incorporates small amounts of carbon dioxide into its crystal structure. Because soil gas carbon dioxide pressures and isotopic compositions are heavily influenced by the oxidation of organic matter with a strongly negative $\delta^{13}C$ value (e.g., -26), there is a gradient in isotopic composition and in the partial pressure of carbon dioxide that develops within the soil profile. This enables the use of minerals such as goethite or carbonate (soil nodules) in paleosols to make a coarse estimate of ancient atmospheric carbon dioxide concentrations. With the exception of one point in the Triassic, the goethite data seem to be in agreement with the forward model estimates of the partial pressure of carbon dioxide in the atmosphere.

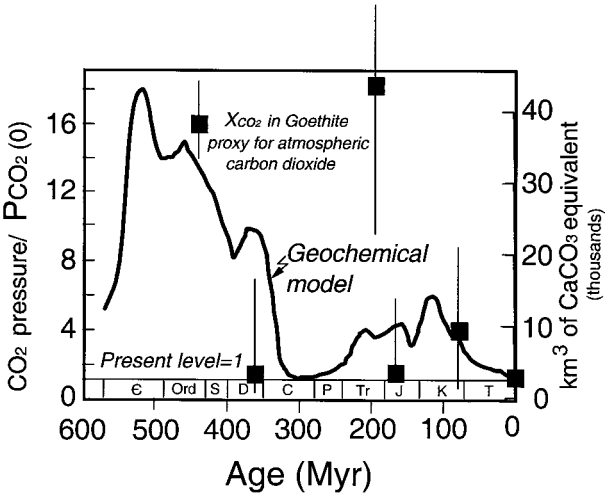


FIGURE 13 Forward geochemical models examine the mass distribution of elements within sedimentary rocks and attempt to document plausible rates of geologic processes to make backwards predictions of important chemical and isotopic parameters. One class of models develops secular change curves for the concentration of atmospheric carbon dioxide, an important greenhouse gas. Shown against one of these models are the few estimates of ancient atmospheric carbon dioxide concentration based upon the isotopic composition of the mineral goethite. The limited data set suggests that stable isotopes have an important role in the documentation of past global change and provide a geologic perspective on current events.

2. Anthropogenic Forcing

In terms of the inferred long-term variability of atmospheric carbon dioxide concentrations, the current icehouse minimum concentrations measured to date (<200 ppmV) may be geologic minimums. Maximum interglacial concentrations have hovered around 300 ppmV carbon dioxide, much less than inferred geologic maximums for the Phanerozoic. Work at the Mauna Loa Observatory (Fig. 14) clearly shows that carbon dioxide in the atmosphere is growing exponentially along with human population. The isotopic composition is shifting to more negative $\delta^{13}C$ values consistent with industrial and agricultural activities of humans.

Clearly unless the rate anthropogenic CO_2 emissions are slowed, the atmospheric concentrations of carbon dioxide will rise to levels not seen for 10s of millions of years and probably not since the last major global warm period. (Figure 15) shows the rates of carbon dioxide cycling for various processes scaled to a parameter that is easily tied to long term geologic rates, cubic kilometers of calcium carbonate processed. *Homo sapiens* is cycling carbon at a geologically unprecedented rate. More than 20 km^3/yr of calcium carbonate equivalent cycles into the atmosphere by fossil fuel burning and land use activity. This rate is only exceeded by the biologic fast carbon cycling rate (respiration and photosynthesis are in rough balance each at $\approx$300 km^3/yr) and the rate of carbon

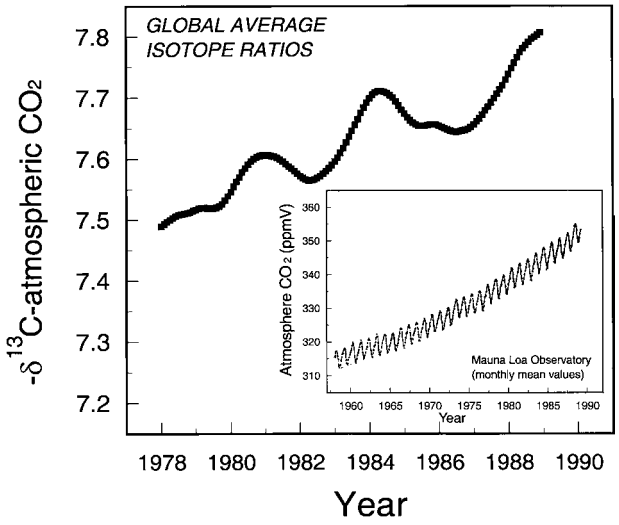


FIGURE 14 Data on the $\delta^{13}C$ value of atmospheric carbon dioxide for the decade starting in 1978. This record, along with the exponential growth of atmospheric carbon dioxide (inset), indicate that the carbon dioxide increase is driven, in large part, by the burning of fossil fuel and deforestation for alternative land use. Typical organic matter and fossil fuel have strongly negative $\delta^{13}C$ values so that their combustion pushes the atmosphere towards more negative values (up on the figure scale).

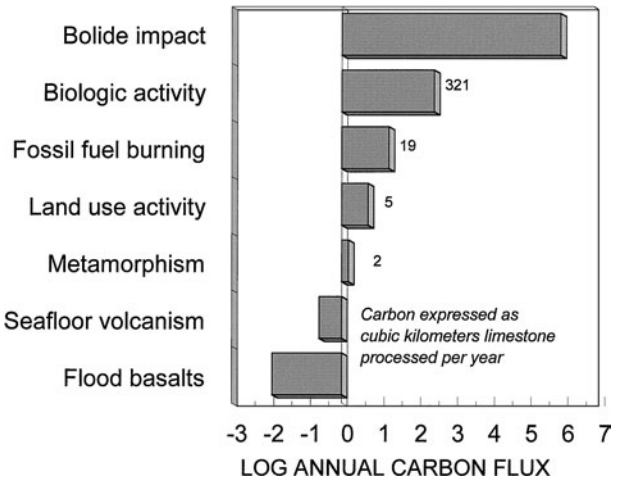


FIGURE 15 Carbon fluxes showing the range of geologic transfers of carbon. The rates are reported in terms of cubic kilometers of limestone equivalents transferred. Long-term geologic rates run at cubic kilometers of carbonate equivalent per year. The short-term carbon cycle processes CO_2 at the rate several hundred cubic kilometers of carbonate equivalent. A bolide impact, such as an event like the Cretaceous-Tertiary boundary, could potentially generate fluxes that would be very catastrophic. Note that on this scale the fluxes produced by *Homo sapiens* sit at geologically unprecedented levels. In Fig. 12, extend the left-hand scale for curve a to over 500 ppmV (i.e., expand the scale by a factor of 6). Note that this is roughly the targeted value for atmospheric carbon dioxide levels by proposed international treaties (Kyoto protocols).

cycling induced by a large impact with an extraterrestrial object.

IV. CONCLUSIONS

Measurements of stable isotopic ratios on common materials from the atmosphere, hydrosphere, lithosphere, and asthenosphere of the Earth document important fluxes between these interacting reservoirs. The active tectonic regime of the Earth coupled with the existence of water at its surface in all forms (ice, liquid, and vapor) induces stable isotopic heterogeneity into the lithosphere of the Earth. Plate tectonics ensures that there are important exchanges of material between high and low temperature reservoirs (e.g., midocean ridges) and provides a mechanism of returning surface material to the mantle of the Earth (e.g., subduction zones). The characteristic times of these long-term exchanges between sources and sinks are short compared to the age of the Earth so that virtually all of the major cycles operate near quasi-steady-state conditions. The chemical and isotopic systems never reach true equilibrium or absolute steady state because the boundary conditions imposed by changes in plate tectonic rates vary slowly over geologic time. As a result, the Earth truly evolves.

SEE ALSO THE FOLLOWING ARTICLES

CONTINENTAL CRUST • EARTH'S MANTLE • HYDROGEOLOGY • IGNEOUS GEOLOGY • MANTLE CONVECTION AND PLUMES • PALYNOLOGY • PLATE TECTONICS • THERMODYNAMICS • STRESS IN THE EARTH'S LITHOSPHERE

BIBLIOGRAPHY

Criss, R. E. (1999). "Principles of Stable Isotope Distribution," Oxford Univ. Press, New York.

Frakes, L. A., Francis, J. E., and Syktus, J. L. (1992). "Climate Modes of the Phanerozoic: The History the Earth's Climate over the Past 600 Million Years," Cambridge Univ. Press, Cambridge.

Gray, D. R., Gregory, R. T., and Durney, D. W. (1991). "Rock-buffered fluid-rock interaction in deformed quartz-rich turbidites sequences, eastern Australia," *J. Geophys. Res.* **96,** 19,681–19,704.

Gregory, R. T. (1991). Oxygen isotope history of seawater revisited: timescales for boundary event changes in the oxygen isotope composition of seawater (H. P. Taylor, Jr., J. R. O'Neil, and I. R. Kaplan, eds.), Volume 3, pp. 65–76. Stable Isotope Geochemistry: A Tribute to Samuel Epstein, Geochemical Society Special Publication.

Hoefs, J. (1997). "Stable Isotope Geochemistry," 4th edition, Spring Verlag, Berlin Heidelberg.

Holser, W. T., Schidlowski, M., Mackenzie, F. T., and Maynard, J. B. (1988). Biogeochemical cycles of carbon and sulfur (C. B. Gregor, R. M. Garrels, F. T. Mackenzie, and J. B. Maynard, eds.), pp. 105–174.

Chemical Cycles in the Evolution of the Earth, Wiley Interscience, New York.

Keeling, C. D., Bacastow, R. B., Carter, A. F., Piper, S. C., Whorf, P., Heimann, M., Mook, W. G., and Roeloffzen (1989). "A three dimensional model of atmospheric CO_2 transport based on observed winds: 1. Analysis of observational data," *Am. Geophys. Union Geophys. Monogr.* **55,** 165–235.

Kyser, T. K. (ed.) (1987). Stable Isotope Geochemistry of Low Temperature Fluids, Mineralogical Association of Canada Short Course, Vol. 17, Toronto.

Petit, J. R., Jouzel, J., Raynaud, D., Barkov, N. I., Barnola, J.-M., Basile, I., Benders, M., Chappellanaz, J., Davis, M., Delaygue, G., Delmotte, M., Kotlyakov, V. M., Legrand, M., Lipenkov, V. Y., Lorius, C., Pepin, L., Ritz, C., Saltzman, E., and Stievenand, M. (1999). "Climate and atlmospheric history of the past 420,000 years from the Vostok ice core, Antarctica," *Nature* **399,** 429–436.

Valley, J. W., Taylor, H. P., Jr., and O'Neil, J. R. (eds.) (1986). "Stable isotopes in high temperature geologic processes," *Mineralogical Soc. Am. Rev. Mineral.* **16**.

Yapp, C. (2001). "Rusty relics of Earth history: iron (III) oxides, isotopes and surficial environments," *Annu. Rev. Earth Planet. Sci.* **29,** 165–199.

Statistical Mechanics

W. A. Wassam, Jr.

El Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional

GLOSSARY

Ensemble A virtual collection of a large number of independent systems, each of which possesses identical thermodynamic properties.

Ensemble average Average value of a property over all members of an ensemble.

Entropy Measure of the statistical uncertainty in the macroscopic state of a system.

Equilibrium Macroscopic state of a system for which the properties are constant over the time scale of measurements and independent of the starting time of the measurements.

Ergodic hypothesis Idea that an isolated system of fixed volume, energy, and particle numbers spends an equal amount of time, over a long time, in each of the accessible states of the system.

Fluctuation–dissipation relation Mathematical relationship connecting the dissipative behavior of a system to spontaneous equilibrium fluctuations.

Maximum entropy principle Principle requiring the probability density or density operator used to describe the macroscopic state of a system to be such that it represents all available information defining the macroscopic state and corresponds to the maximum Gibbs entropy distribution consistent with this information.

Principle of equal a priori probabilities Idea that all accessible states are equally probable for an isolated system with fixed volume, energy, and particle numbers.

Response function Time correlation function characterizing the response of a system at a given space-time point to a disturbance at another space-time point.

Susceptibility Frequency-dependent quantity used to gauge the response of a system to a frequency component of an applied external field.

Time correlation function Time-dependent quantity characterizing the correlation between dynamical events occurring at different times in a system.

STATISTICAL MECHANICS is a branch of physics that utilizes statistical methods in conjunction with the basic

principles of classical and quantum mechanics to provide a theoretical framework for understanding equilibrium and nonequilibrium properties of physical systems. Usually statistical mechanics is divided into two parts, namely, equilibrium and nonequilibrium statistical mechanics. Equilibrium statistical mechanics is concerned with the time-independent properties of systems, whereas nonequilibrium statistical mechanics is concerned with the time evolution of physical systems and their time-dependent properties. The principal goals of statistical mechanics include (i) the construction of mathematical relationships connecting different properties of a system, (ii) the establishment of formal relations connecting the properties of a system to the interactions between the constituent particles, (iii) the development of tools that enable one to determine properties of systems without resorting to the task of solving global equations of motion, and (iv) the formulation of a theory of nonequilibrium processes that automatically includes equilibrium thermodynamics as a special limiting case.

I. CONCEPTS OF EQUILIBRIUM STATISTICAL MECHANICS

A. Notion of Equilibrium

The macroscopic state of a system can be characterized in terms of its properties, i.e., the external parameters defining the system and the values of experimentally measurable attributes of the system. These properties may include volume, pressure, stress, magnetization, polarization, number of particles, etc.

In performing measurements of the properties of a system, experimentalists often find that the values of the measured properties are constant over the time scale $\Delta t = t - t_0$ of their measurements and independent of the starting time t_0 of the measurements. When such results are obtained, the system is said to be in the state of equilibrium. Of course, we know from microscopic considerations that the properties of a system that are not fixed by external restraints must be a function of time. Thus, the appearance of equilibrium must be due to the smoothing out of fluctuations on the time scale of macroscopic measurements.

B. Concept of Ensembles

In order to describe the macroscopic state of a system, we introduce the idea of a representative ensemble. By ensemble, we mean a virtual collection of a large number $\mathcal{N}$ of independent systems, each of which possesses identical thermodynamic properties. Although all of the members of the ensemble are identical from a thermodynamic point of view, they can differ on a microscopic scale. The ensemble corresponding to a given system is said to be a representation of the macroscopic state of the system.

The specification of the representative ensemble of a given system requires us to enumerate the external parameters defining the system and the integrals of motion characterizing the system. According to Gibbs, the macroscopic state corresponding to equilibrium depends only on single-valued additive integrals of motion. By additive, we mean that the integrals of motion are additively composed of the integrals of motion of the subsystems comprising the system.

Consider, for example, a system of fixed energy E, number of particles N, and volume V. This system is energetically isolated and closed to particle transfer across its boundaries. For such a system, there are four integrals of motion, namely, the total energy E, the total linear momentum $\vec{P}$, the total angular momentum $\vec{L}$, and the total number of particles N. Thus, we expect the macroscopic state of the system to depend on E, $\vec{P}$, $\vec{L}$, N, and V. If the overall system is motionless, $|\vec{P}| = |\vec{L}| = 0$. For this case, the macroscopic state depends only on E, N, and V. This macroscopic state is said to be represented by the microcanonical ensemble.

The ensemble average of a property of a system is defined as the average value of the property over all members of the ensemble. In usual formulations of equilibrium statistical mechanics, it is postulated that the ensemble average of a property of a system is equivalent to the corresponding thermodynamic property. (For example, the ensemble average of the energy is equivalent to the thermodynamic energy or the internal energy.) In adopting this postulate, one is asserting that the long-time average of a property of a system in equilibrium is equal to the ensemble average of the property in the limit $\mathcal{N} \to \infty$, provided the members of the ensemble replicate the thermodynamic state of the actual system.

C. Macroscopic State of Classical Systems

The dynamical state of a classical particle l is defined by its coordinate (position) $\vec{q}_l$ and momentum $\vec{p}_l$. The state of a collection of N particles is defined by specifying the collective coordinate $\vec{q}^N = (\vec{q}_1, \ldots, \vec{q}_N)$ and collective momentum $\vec{p}^N = (\vec{p}_1, \ldots, \vec{p}_N)$. Alternatively, we can define the state of an N-particle system by specifying the phase point $\vec{\Gamma}^N = (\vec{p}^N, \vec{q}^N)$ in a $6N$-dimensional space called phase space.

The ensemble representing the macroscopic state of a classical system is described by an object called the probability density $\rho(\vec{p}^N, \vec{q}^N)$, which is defined in such a way that

$$dP(\vec{p}^N, \vec{q}^N) = d^{3N}p d^{3N}q \rho(\vec{p}^N, \vec{q}^N) \tag{1}$$

represents the probability of finding the system in the phase space volume element $d^{3N}p\, d^{3N}q$ in the neighborhood of the phase point $(\vec{p}^N, \vec{q}^N)$. The sum of the probablities must be equal to unity. Thus,

$$\int d^{3N}p \int d^{3N}q \rho(\vec{p}^N, \vec{q}^N) = 1. \tag{2}$$

Since classical mechanics corresponds to a limit of quantum mechanics, another normalization of the probability density is usually adopted. For a system of N identical particles, this normalization is written

$$\int d^{6N}\Gamma\, \rho(\vec{p}^N, \vec{q}^N) = 1, \tag{3}$$

where

$$d^{6N}\Gamma = \left(\frac{1}{N!h^{3N}}\right) d^{3N}p d^{3N}q \tag{4}$$

is a dimensionless volume element in the $6N$-dimensional phase space, with h denoting Planck's constant. The motivation for introducing this normalization is based on the idea that the normalization should reflect Heisenberg's uncertainty principle and the indistinguishability of identical particles in quantum mechanics. The rationale for adopting the above explicit form for $d^{6N}\Gamma$ is discussed below.

In view of Heisenberg's uncertainty principle, there is a minimum volume h^3 in phase space that may be associated with a single particle. This minimum volume is called a phase cell. In the phase space of N particles, the volume of a phase cell is h^{3N}. Thus, h^{3N} is a natural unit of volume in the phase space for an N-particle system.

The indistinguishability of identical particles is accounted for in quantum mechanics by requiring the state of a system to be invariant to the permutation of particle coordinates. This invariance must be preserved in the classical limit. For a quantum system of N identical particles, there are $N!$ configurations (permutations) corresponding to a given state. Thus, there are $N!$ configurations for a given classical state. (Actually, the situation is more complicated than this. Our discussion corresponds to a limiting form of both Bose–Einstein and Fermi–Dirac statistics called classical or Boltzmann statistics, which is valid only when the number of states available to a system is very large compared with N.)

Introducing the volume unit h^{3N} and the indistinguishability of identical particles, we arrive at the normalization given by Eq. (3). The integration in Eq. (3) is simply a summation over the different states of the classical system.

For a classical system comprised of m different kinds of particles, we replace $d^{6N}\Gamma$ by

$$d^{6N}\Gamma = \prod_{j=1}^{m} d^{6N_j}\Gamma_j, \tag{5}$$

where

$$d^{6N_j}\Gamma_j = \left(\frac{1}{N_j!h^{3N_j}}\right) d^{3N_j}p_j d^{3N_j}q_j. \tag{6}$$

Hereafter, this notation should be understood.

Given the probability density $\rho(\vec{p}^N, \vec{q}^N)$, we can compute the ensemble average or equivalently the average value $\langle O\rangle$ of any classical dynamical variable $O(\vec{p}^N, \vec{q}^N)$ by using the relation

$$\langle O\rangle = \int d^{6N}\Gamma\, \rho(\vec{p}^N, \vec{q}^N) O(\vec{p}^N, \vec{q}^N). \tag{7}$$

For example, the average energy of a N-particle system is given by

$$\langle \mathcal{H}\rangle = \int d^{6N}\Gamma\, \rho(\vec{p}^N, \vec{q}^N) \mathcal{H}(\vec{p}^N, \vec{q}^N), \tag{8}$$

where $\mathcal{H}(\vec{p}^N, \vec{q}^N)$ is the classical Hamiltonian of the system.

The Gibbs entropy associated with the ensemble described by $\rho(\vec{p}^N, \vec{q}^N)$ is defined by

$$S = -k_{\mathrm{B}} \int d^{6N}\Gamma\, \rho(\vec{p}^N, \vec{q}^N) \ln \rho(\vec{p}^N, \vec{q}^N), \tag{9}$$

where k_{B} is Boltzmann's constant. This definition may be viewed as another postulate of equilibrium statistical mechanics for the case of classical systems.

D. Macroscopic State of Quantum Systems

In quantum mechanics, the state of an N-particle system is described by a state vector $|\psi\rangle$ in Hilbert space. The wave function $\psi(\vec{q}^N)$ corresponding to the state $|\psi\rangle$ is defined by $\psi(\vec{q}^N) = \langle \vec{q}^N \mid \psi\rangle$, where $\langle \vec{q}^N|$ is the adjoint of the collective coordinate state vector $|\vec{q}^N\rangle$, with $\vec{q}^N = (\vec{q}_1, \ldots, \vec{q}_N)$ denoting the collective coordinate for the N particles.

The eigenstates $\{|\psi_m\rangle\}$ and eigenvalues $\{E_m\}$ of the quantum Hamiltonian $\hat{\mathcal{H}}$ for an N-particle system are obtained by solving the equation

$$\hat{\mathcal{H}}|\psi_m\rangle = E_m|\psi_m\rangle. \tag{10}$$

For a system of N identical particles of mass m interacting through a pairwise and centrally symmetric potential, the Hamiltonian $\hat{\mathcal{H}}$ can be written

$$\hat{\mathcal{H}} = \sum_{i=1}^{N} \frac{\hat{\vec{p}}_i \cdot \hat{\vec{p}}_i}{2m} + \frac{1}{2}{\sum_{i,j=1}^{N}}' \hat{U}_{ij}(|\hat{\vec{q}}_i - \hat{\vec{q}}_j|), \tag{11}$$

where $\hat{\vec{p}}_i$ and $\hat{\vec{q}}_i$, respectively, denote the momentum and coordinate vector operators for particle i, and $\hat{U}_{ij}$ is the interaction between particles i and j.

Equation (10) assumes the following form in the coordinate representation:

$$\mathcal{H}(\vec{q}^N)\psi_m(\vec{q}^N) = E_m\psi_m(\vec{q}^N), \tag{12}$$

where $\psi_m(\vec{q}^N) = \langle \vec{q}^N|\psi_m\rangle$ is an eigenfunction of the differential operator $\mathcal{H}(\vec{q}^N)$ corresponding to the quantum Hamiltonian $\hat{\mathcal{H}}$ in the coordinate representation. For the case of Eq. (11), the explicit form of $\mathcal{H}(\vec{q}^N)$ is given by

$$\mathcal{H}(\vec{q}^N) = -\sum_{i=1}^{N}\frac{\hbar^2}{2m}\nabla_{q_i}^2 + \frac{1}{2}\sum_{i,j=1}^{N}{}'U_{ij}(|\vec{q}_i - \vec{q}_j|), \tag{13}$$

where $\nabla_{q_i}^2$ is a Laplacian operator associated with coordinates of particle i.

In order to provide a proper description of the N-particle system, it is necessary to specify its spin state as well as energy state. Assuming that $\mathcal{H}(\vec{q}^N)$ does not mix the spin states, we should write Eq. (12) as

$$\mathcal{H}(\vec{q}^N)\psi_{m\sigma}(\vec{q}^N) = E_{m\sigma}\psi_{m\sigma}(\vec{q}^N), \tag{14}$$

where the index σ has been introduced to specify the spin. Not all the eigenfunctions of Eq. (14) are permissible. Only those satisfying certain symmetry requirements are allowed. More specifically, the eigenfunctions for a system of particles with zero or integral (half-integral) spin, in multiples of $\hbar = h/2\pi$, are required to be symmetric (antisymmetric) with respect to the interchange of particle coordinates. The particles are said to obey Bose–Einstein statistics for the case of zero or integral spin. Otherwise, the particles obey Fermi–Dirac statistics.

One can partition the Hamiltonian $\mathcal{H}(\vec{q}^N)$ as follows:

$$\mathcal{H}(\vec{q}^N) = \mathcal{H}^{(0)}(\vec{q}^N) + U(\vec{q}^N), \tag{15}$$

where

$$\mathcal{H}^{(0)}(\vec{q}^N) = -\sum_{i=1}^{N}\frac{\hbar^2}{2m}\nabla_{q_i}^2 \tag{16}$$

describes the free motion of the particles and

$$U(\vec{q}^N) = \frac{1}{2}\sum_{i,j=1}^{N}{}'U_{ij}(|\vec{q}_i - \vec{q}_j|) \tag{17}$$

is the interaction between the particles.

The eigenfunctions $\{\psi_{m\sigma}(\vec{q}^N)\}$ of the Hamiltonian $\mathcal{H}(\vec{q}^N)$ can be constructed by taking linear combinations of the eigenfunctions $\{\phi_{l\sigma}^{(0)}(\vec{q}^N)\}$ of $\mathcal{H}^{(0)}(\vec{q}^N)$:

$$\psi_{m\sigma}(\vec{q}^N) = \sum_{l} C_{m\sigma,l\sigma}\phi_{l\sigma}^{(0)}(\vec{q}^N). \tag{18}$$

From this equation, we see that the eigenfunctions $\{\psi_{m\sigma}(\vec{q}^N)\}$ of $\mathcal{H}(\vec{q}^N)$ possess the proper symmetry when the eigenfunctions $\{\phi_{l\sigma}^{(0)}(\vec{q}^N)\}$ of $\mathcal{H}^{(0)}(\vec{q}^N)$ satisfy this requirement.

For the case of Bose–Einstein statistics, the eigenfunctions $\{\psi_{m\sigma}(\vec{q}^N)\}$ are required to be symmetric with respect to particle interchange. Thus, the eigenfunctions $\{\phi_{l\sigma}^{(0)}(\vec{q}^N)\}$ must be symmetric. These symmetric eigenfunctions can be written

$$\phi_{l\sigma}^{(0)}(\vec{q}^N) = \frac{1}{\sqrt{N!}}\sum_{\nu}\mathcal{P}_\nu\big[\chi_{l_1\sigma_1}(\vec{q}_1)\cdots\chi_{l_N\sigma_N}(\vec{q}_N)\big], \tag{19}$$

where $\mathcal{P}_\nu$ is a permutation operator that permutes particle coordinates, the index ν runs over the $N!$ possible permutations, and $\chi_{l_l\sigma_l}(\vec{q}_l)$ denotes a one-particle state occupied by particle l. The one-particle states $\{\chi_{l_l\sigma_l}(\vec{q}_l)\}$ are the solutions of

$$-(\hbar^2/2m)\nabla_{q_l}^2\chi_{l_l\sigma_l}(\vec{q}_l) = E_{l_l\sigma_l}^{(0)}\chi_{l_l\sigma_l}(\vec{q}_l). \tag{20}$$

For the case of Fermi–Dirac statistics, the eigenfunctions $\{\psi_{m\sigma}(\vec{q}^N)\}$ are required to be antisymmetric with respect to particle exchange. Thus, the eigenfunctions $\{\phi_{l\sigma}^{(0)}(\vec{q}^N)\}$ must be antisymmetric. These antisymmetric eigenfunctions can be written

$$\phi_{l\sigma}^{(0)}(\vec{q}^N) = \frac{1}{\sqrt{N!}}\sum_{\nu}\epsilon_\nu\mathcal{P}_\nu\big[\chi_{l_1\sigma_1}(\vec{q}_1)\cdots\chi_{l_N\sigma_N}(\vec{q}_N)\big], \tag{21}$$

where $\epsilon_\nu = +1/-1$ for an even/odd permutation of particle coordinates.

In quantum mechanics, dynamical variables are represented by linear Hermitian operators $\hat{O}$ that operate on state vectors in Hilbert space. The spectra of these operators determine possible values of the physical quantities that they represent. Unlike classical systems, specifying the state $|\psi\rangle$ of a quantum system does not necessarily imply exact knowledge of the value of a dynamical variable. Only for cases in which the system is in an eigenstate of a dynamical variable will the knowledge of that state $|\psi\rangle$ provide an exact value. Otherwise, we can only determine the quantum average of the dynamical variable.

The quantum average $\langle\hat{O}\rangle_\psi$ of the dynamical variable $\hat{O}$ is given by

$$\langle\hat{O}\rangle_\psi = \langle\psi|\hat{O}|\psi\rangle \tag{22a}$$

$$= \int d^{3N}q\,\psi^*(\vec{q}^N)O(\vec{q}^N)\psi(\vec{q}^N), \tag{22b}$$

where $O(\vec{q}^N)$ is defined by

$$\langle\vec{q}^N|\hat{O}|\vec{q}'^N\rangle = O(\vec{q}^N)\delta(\vec{q}^N - \vec{q}'^N), \tag{23}$$

with $\delta(\vec{q}^N - \vec{q}'^N)$ denoting a Dirac delta function. For the sake of simplicity, we have neglected any possible spin dependence of the quantum average.

The quantum average $\langle \hat{O} \rangle_\psi$ of the dynamical variable $\hat{O}$ is actually a conditional average, i.e., it is conditional on the fact that the system of interest has been prepared in the state $|\psi\rangle$. In general, we lack sufficient information to specify the state of a system. Thus, additional probabilistic concepts must be introduced that are not inherent in the formal structure of quantum mechanics.

Suppose we know the set of probabilties $\{P(\psi)\}$ for observing a system in the different states $\{|\psi\rangle\}$. Then the average value of the dynamical variable $\hat{O}$ is given by the weighted quantum average

$$\langle \hat{O} \rangle = \sum_\psi P(\psi) \langle \hat{O} \rangle_\psi . \tag{24}$$

In order to compute the weighted quantum average, we need something to provide us with information about the probabilities$\{P(\psi)\}$. This task is fulfilled by a Hermitian operator $\hat{\rho}$ called the density operator. The density operator $\hat{\rho}$ is the quantum analogue of the probability density ρ for classical systems. Thus, we say that the macroscopic state of a quantum system is described by the density operator $\hat{\rho}$.

The density operator $\hat{\rho}$ possesses the following properties: (i) It is a Hermitian operator, i.e., $\hat{\rho} = \hat{\rho}^\dagger$. (ii) The trace of $\hat{\rho}$ is unity, i.e., $\mathrm{Tr}\, \hat{\rho} = 1$. (iii) The diagonal matrix elements of $\hat{\rho}$ in any representation are non-negative, i.e., $\langle \alpha | \hat{\rho} | \alpha \rangle \geq 0$. (iv) The probability of finding a system in some state $|\alpha\rangle$ is given by the diagonal matrix element $\rho(\alpha, \alpha) = \langle \alpha | \hat{\rho} | \alpha \rangle$.

The eigenvectors $\{|\psi\rangle\}$ and eigenvalues $\{P(\psi)\}$ of $\hat{\rho}$ are determined by solving the equation

$$\hat{\rho} |\psi\rangle = P(\psi) |\psi\rangle, \tag{25}$$

where

$$P(\psi) = \langle \psi | \hat{\rho} | \psi \rangle \tag{26}$$

is the probability of finding the system in the state $|\psi\rangle$. In the ψ-representation, we write

$$\hat{\rho} = \sum_{\psi'} |\psi'\rangle P(\psi') \langle \psi' |. \tag{27}$$

Although the density operator $\hat{\rho}$ is diagonal in the ψ-representation, this may not be the case in any other representation. In other representations, there may be non-vanishing off-diagonal matrix elements. For example, we have in the coordinate representation

$$\rho(\vec{q}^N, \vec{q}'^N) = \sum_\psi \psi(\vec{q}^N) P(\psi) \psi^*(\vec{q}'^N), \tag{28}$$

where the wave function $\psi(\vec{q}^N) = \langle \vec{q}^N | \psi \rangle$ is the probability amplitude for finding the system with the collective coordinate $\vec{q}^N$, given that the system is in the state $|\psi\rangle$. Clearly, $\rho(\vec{q}^N, \vec{q}^N)$ gives us the probability density for finding the system with the collective coordinate $\vec{q}^N$. The off-diagonal matrix element $\rho(\vec{q}^N, \vec{q}'^N)$ tells us something about the phase coherence between the collective coordinate states $|\vec{q}^N\rangle$ and $|\vec{q}'^N\rangle$, i.e., the interference between the probability amplitudes $\psi(\vec{q}^N)$ and $\psi^*(\vec{q}'^N)$.

The macroscopic state described by the density operator $\hat{\rho}$ may be either a pure state or a mixed state. This classification scheme is based on the character of the trace $\mathrm{Tr}\, \hat{\rho}^2$. More specifically, $\mathrm{Tr}\, \hat{\rho}^2 = 1$ for pure states and $\mathrm{Tr}\, \hat{\rho}^2 < 1$ for mixed states.

The criterion for a pure state can be satisfied only if the eigenvalues of $\hat{\rho}$ satisfy the condition $P(\psi') = \delta_{\psi', \psi}$. This implies that the system is in some quantum state $|\psi\rangle$. Thus, the density operator $\hat{\rho}$ corresponding to a pure state is of the form

$$\hat{\rho} = |\psi\rangle\langle\psi|. \tag{29}$$

If the density operator for the system is not of this form, the criterion for a mixed state will be fulfilled. For this case, the density operator $\hat{\rho}$ is of the form of Eq. (27) with a spread in the eigenvalues $\{P(\psi')\}$.

Given the density operator $\hat{\rho}$, we can compute the ensemble average or equivalently the average value $\langle \hat{O} \rangle$ of any quantum dynamical variable $\hat{O}$ by using the relations

$$\langle \hat{O} \rangle = \mathrm{Tr}\, \hat{\rho} \hat{O} \tag{30a}$$

$$= \sum_\psi P(\psi) \langle \hat{O} \rangle_\psi . \tag{30b}$$

Equation (30a) is more useful than Eq. (30b). The latter equation requires the determination of the eigenvectors of $\hat{\rho}$. This is by no means a trivial task. Since the trace in Eq. (30a) is representation independent, it can be evaluated by employing any convenient representation satisfying the proper symmetry requirements and any other restraints placed on the system. Note that Eq. (30b) assumes the form of a quantum average when the system has been prepared in a pure state.

The Gibbs entropy associated with the ensemble described by $\hat{\rho}$ is defined by

$$S = -k_\mathrm{B}\, \mathrm{Tr}\, \hat{\rho} \ln \hat{\rho}. \tag{31}$$

This definition is the quantum analogue of the definition given for classical systems. Both the quantum and classical forms for the entropy S involve summations over the accessible states of a system.

In the representation for which $\hat{\rho}$ is diagonal, the Gibbs entropy S can be written

$$S = -k_\mathrm{B} \sum_\psi P(\psi) \ln P(\psi). \tag{32}$$

This equation reveals that the entropy S vanishes when a system is prepared in a pure state. This is due to the lack

of any statistical indeterminancy in the state of the system. For the case of a mixed state, the entropy is nonzero. This reflects the fact that there is some statistical indeterminancy in the description of the system.

II. GIBBSIAN ENSEMBLES AND THERMODYNAMICS

A. Microcanonical Ensemble

1. Classical Systems

Consider an energetically isolated (adiabatic) system composed of a fixed number of particles N (closed) and enclosed in a motionless container of fixed volume V. The energy of the system, given by the classical Hamiltonian $\mathcal{H}(\vec{p}^N, \vec{q}^N)$, lies in an energy shell of width ΔE, i.e., $E \leq \mathcal{H}(\vec{p}^N, \vec{q}^N) \leq E + \Delta E$. The macroscopic state of this system is called a microcanonical ensemble.

For a multicomponent system, we shall use the symbol N to represent the collection of particle numbers $\{N_j\}$ for the components of the system. Hereafter, this notation should be understood.

The probability density ρ for a classical microcanonical ensemble is given by

$$\rho(\vec{p}^N, \vec{q}^N) = \Omega^{-1}(E, V, N; \Delta E) \tag{33}$$

for the phase points lying inside the energy shell $E \leq \mathcal{H}(\vec{p}^N, \vec{q}^N) \leq E + \Delta E$. Outside the energy shell, $\rho(\vec{p}^N, \vec{q}^N)$ vanishes. In the above, $\Omega(E, V, N; \Delta E)$ is the number of states inside the energy shell, i.e.,

$$\Omega(E, V, N; \Delta E) = \int_{E \leq \mathcal{H}(\vec{p}^N, \vec{q}^N) \leq E + \Delta E} d^{6N}\Gamma. \tag{34}$$

According to Eq. (33), all states inside the energy shell are equally probable. In other words, the microcanonical ensemble represents a uniform distribution over the phase points belonging to the energy shell. Such a distribution is often referred to as a mathematical statement of the principle of equal a priori probabilities. According to this principle, the system spends an equal amount of time, over a long time period, in each of the available classical states. This statement is called the ergodic hypothesis.

The Gibbs entropy of the distribution described by Eq. (33) is given by

$$S = k_B \ln \Omega(E, V, N; \Delta E). \tag{35}$$

This result corresponds to Boltzmann's definition of entropy. Obviously, the entropy S increases as the number of accessible states increases. This behavior of the Boltzmann entropy has led to such qualitative definitions of entropy as "measure of randomness" and "measure of disorder." The microcanonical distribution is the distribution of maximum Gibbs entropy consistent with knowing nothing about the system except the specified values of E, V, and N.

The entropy S of the microcanonical ensemble depends only on the variables E, V, and N. Thus, the differential dS can be written

$$dS = \left(\frac{\partial S}{\partial E}\right)_{V,N} dE + \left(\frac{\partial S}{\partial V}\right)_{E,N} dV + \sum_l \left(\frac{\partial S}{\partial N_l}\right)_{E,V,N'} dN_l, \tag{36}$$

where the index l runs over the components of the system,

$$(\partial S/\partial E)_{V,N} = k_B[\partial \ln \Omega(E, V, N; \Delta E)/\partial E]_{V,N}, \tag{37}$$

$$(\partial S/\partial V)_{E,N} = k_B[\partial \ln \Omega(E, V, N; \Delta E)/\partial V]_{E,N}, \tag{38}$$

and

$$\left(\frac{\partial S}{\partial N_l}\right)_{E,V,N'} = k_B[\partial \ln \Omega(E, V, N; \Delta E)/\partial N_l]_{E,V,N'}. \tag{39}$$

In Eqs. (36) and (39), we use the symbol N' to indicate that the particle numbers $N' = \{N_j; j \neq l\}$ are fixed. Hereafter, this notation should be understood.

Entropy S is the thermodynamic characteristic function for the variables E, V, and N:

$$dS = \frac{1}{T} dE + \frac{P}{T} dV - \sum_l \frac{\mu_l}{T} dN_l, \tag{40}$$

where

$$\frac{1}{T} = \left(\frac{\partial S}{\partial E}\right)_{V,N}, \tag{41}$$

$$\frac{P}{T} = \left(\frac{\partial S}{\partial V}\right)_{E,N}, \tag{42}$$

and

$$\frac{\mu_l}{T} = -\left(\frac{\partial S}{\partial N_l}\right)_{E,V,N'}. \tag{43}$$

In the above, E is the internal energy, T the temperature, P the pressure, V the volume, μ_l the chemical potential of component l, and N_l the number of particles of component l.

Assuming the Gibbs entropy for the microcanonical ensemble is the same as the thermodynamic entropy, we can use the above results to construct the following statistical mechanical expressions for the thermodynamic quantities T, P, and μ_l:

$$\frac{1}{T} = k_B[\partial \ln \Omega(E, V, N; \Delta E)/\partial E]_{V,N}, \quad (44)$$

$$\frac{P}{T} = k_B[\partial \ln \Omega(E, V, N; \Delta E)/\partial V]_{E,N}, \quad (45)$$

and

$$\frac{\mu_l}{T} = -k_B[\partial \ln \Omega(E, V, N; \Delta E)/\partial N_l]_{E,V,N'}. \quad (46)$$

In principle, these results can be employed to compute T, P, and μ_l. In practice, however, this is extremely difficult because the number of states $\Omega(E, V, N; \Delta E)$ as a function of E, V, and N is usually not available.

2. Quantum Systems

As for classical systems, the microcanonical ensemble is used to characterize the macroscopic state of energetically isolated and closed quantum system of fixed volume. The density operator $\hat{\rho}$ for a quantum microcanonical ensemble is given by

$$\hat{\rho} = \Omega^{-1}(E, V, N; \Delta E)\hat{P}, \quad (47)$$

where $\Omega(E, V, N; \Delta E)$ is the number of eigenstates $\{|\psi_l\rangle\}$ of the system Hamiltonian $\hat{\mathcal{H}}$ lying in the energy shell $[E, E + \Delta E]$ of width ΔE, and $\hat{P}$ is a projection operator defined by

$$\hat{P} = \sum_{l \epsilon \Omega} |\psi_l\rangle\langle\psi_l|, \quad (48)$$

with the index l enumerating the eigenstates belonging to the energy shell Ω.

The probability $P(\psi_m)$ of finding a system described by the microcanonical density operator $\hat{\rho}$ in a given eigenstate $|\psi_m\rangle$ of $\hat{\mathcal{H}}$ is given by the diagonal matrix element $\langle\psi_m|\hat{\rho}|\psi_m\rangle$. For an eigenstate $|\psi_m\rangle$ lying in the energy shell $[E, E + \Delta E]$, $P(\psi_m)$ is given by

$$P(\psi_m) = \Omega^{-1}(E, V, N; \Delta E). \quad (49)$$

Otherwise, $P(\psi_m)$ vanishes. From these results, we see that the quantum microcanonical ensemble describes a uniform distribution over the eigenstates of the system Hamiltonian $\hat{\mathcal{H}}$ lying inside the energy shell $[E, E + \Delta E]$. Outside the energy shell, the distribution vanishes.

The Gibbs entropy of the quantum microcanonical ensemble is identical in form to the Gibbs entropy of the classical microcanonical ensemble, except $\Omega(E, V, N; \Delta E)$ must now be interpreted as the number of eigenstates of the system Hamiltonian lying in the energy shell. As with the microcanonical probability density ρ for classical systems, the microcanonical density operator $\hat{\rho}$ for quantum systems represents the distribution of maximum Gibbs entropy consistent with knowing nothing about the system except the specified values of E, V, and N.

In view of the formal identity of the expressions for the Gibbs entropy of quantum and classical microcanonical ensembles, the statistical mechanical expressions given by Eqs. (44)–(46) also apply to quantum systems. One need only reinterpret the quantity $\Omega(E, V, N; \Delta E)$ appearing in these expressions as the number of eigenstates of the system Hamiltonian $\hat{\mathcal{H}}$ lying in the energy shell $[E, E + \Delta E]$.

B. Canonical Ensemble

1. Classical Systems

Consider a system made up of a fixed number of particles N enclosed in a motionless container of fixed volume V. The temperature of the system is maintained at the temperature T by keeping the system in thermal contact with a large heat bath at the temperature T with which it is able to exchange energy. Such a system is called a closed, isothermal system. The macroscopic state of this system is called the canonical ensemble.

The probability density ρ for a classical canonical ensemble is given by

$$\rho(\vec{p}^N, \vec{q}^N) = Z^{-1}(V, N, \beta) \exp[-\beta \mathcal{H}(\vec{p}^N, \vec{q}^N)], \quad (50)$$

where β is a parameter to be determined and

$$Z(V, N, \beta) = \int d^{6N}\Gamma \exp[-\beta \mathcal{H}(\vec{p}^N, \vec{q}^N)] \quad (51)$$

is the canonical partition function. The canonical partition function $Z(V, N, \beta)$ depends on the three parameters V, N, and β.

The Gibbs entropy of the canonical distribution given by Eq. (50) can be written as

$$S = k_B \beta \langle \mathcal{H} \rangle + k_B \ln Z(V, N, \beta), \quad (52)$$

where $\langle \mathcal{H} \rangle$ denotes the average energy of the system. The canonical distribution is the distribution of maximum Gibbs entropy consistent with the given average energy $\langle \mathcal{H} \rangle$.

Equation (52) can be rearranged to read

$$\langle \mathcal{H} \rangle = -\beta^{-1} \ln Z(V, N, \beta) + (k_B \beta)^{-1} S. \quad (53)$$

This result resembles the thermodynamic relation

$$E = A + TS, \quad (54)$$

where E is the internal energy, A the Helmholtz free energy, T the temperature, and S the entropy.

Assuming the Gibbs entropy of the canonical ensemble is equivalent to the thermodynamic entropy, we have that the average energy $\langle \mathcal{H} \rangle$ is equivalent to the internal energy E, and with $\beta = 1/k_B T$, we can write

$$A = -\beta^{-1} \ln Z(V, N, \beta). \quad (55)$$

This result connects the Helmholtz free energy A to microscopic interactions through the canonical partition function $Z(V, N, \beta)$, which depends on the classical Hamiltonian $\mathcal{H}(\vec{p}^N, \vec{q}^N)$.

The Helmholtz free energy A is the thermodynamic characteristic function for the variables V, N, and T:

$$dA = -S\,dT - P\,dV + \sum_l \mu_l\,dN_l, \tag{56}$$

where

$$S = -\left(\frac{\partial A}{\partial T}\right)_{V,N}, \tag{57}$$

$$P = -\left(\frac{\partial A}{\partial V}\right)_{T,N}, \tag{58}$$

and

$$\mu_l = \left(\frac{\partial A}{\partial N_l}\right)_{V,T,N'}. \tag{59}$$

In the above, S is the entropy, T the temperature, P the pressure, μ_l the chemical potential of component l, and N_l the number of particles of component l.

Utilizing the above results, one can construct the following statistical mechanical relations for the thermodynamic quantities E, S, P, and μ_l:

$$E = k_B T^2[\partial \ln Z(V, N, \beta)/\partial T]_{V,N}, \tag{60}$$

$$S = k_B T[\partial \ln Z(V, N, \beta)/\partial T]_{V,N} + k_B \ln Z(V, N, \beta), \tag{61}$$

$$P = k_B T[\partial \ln Z(V, N, \beta)/\partial V]_{T,N}, \tag{62}$$

and

$$\mu_l = -k_B T[\partial \ln Z(V, N, \beta)/\partial N_l]_{V,T,N'}. \tag{63}$$

An expression for the average energy $\langle \mathcal{H} \rangle$ can be obtained by differentiating $\ln Z(V, N, \beta)$ with respect to β:

$$\langle \mathcal{H} \rangle = -[\partial \ln Z(V, N, \beta)/\partial \beta]_{V,N}. \tag{64}$$

Differentiating $\ln Z(V, N, \beta)$ a second time with respect to β gives us the following expressions for the energy fluctuations in the system:

$$\langle [\mathcal{H} - \langle \mathcal{H} \rangle]^2 \rangle = \langle \mathcal{H}^2 \rangle - \langle \mathcal{H} \rangle^2 \tag{65a}$$

$$= [\partial^2 \ln Z(V, N, \beta)/\partial \beta^2]_{V,N} \tag{65b}$$

$$= -(\partial \langle \mathcal{H} \rangle / \partial \beta)_{V,N}. \tag{65c}$$

The above results enable us to connect the heat capacity $C_V = (\partial \langle \mathcal{H} \rangle / \partial T)_{V,N}$ at a constant volume to energy fluctuations and microscopic interactions through the relations

$$C_V = k_B \beta^2 \langle [\mathcal{H} - \langle \mathcal{H} \rangle]^2 \rangle \tag{66a}$$

$$= k_B \beta^2 [\partial^2 \ln Z(V, N, \beta)/\partial \beta^2]_{V,N}. \tag{66b}$$

2. Quantum Systems

As for classical systems, the canonical ensemble is used to characterize the macroscopic state of a closed, isothermal quantum system. The density operator $\hat{\rho}$ for quantum canonical ensemble is given by

$$\hat{\rho} = Z^{-1}(V, N, \beta) \exp(-\beta \hat{\mathcal{H}}), \tag{67}$$

where V is the volume, $N = \{N_j\}$ the collection of particle numbers, $\beta = 1/k_B T$, $\hat{\mathcal{H}}$ the Hamiltonian, and

$$Z(V, N, \beta) = \mathrm{Tr} \exp(-\beta \hat{\mathcal{H}}) \tag{68}$$

the canonical partition function.

The probability $P(\psi_m)$ of finding a system described by the canonical density operator $\hat{\rho}$ in a given eigenstate $|\psi_m\rangle$ of the system Hamiltonian $\hat{\mathcal{H}}$ is given by the diagonal matrix element $\langle \psi_m | \hat{\rho} | \psi_m \rangle$:

$$P(\psi_m) = Z^{-1}(V, N, \beta) \exp(-\beta E_m), \tag{69}$$

where E_m is the energy of the state $|\psi_m\rangle$. The canonical partition function Z can be written

$$Z(V, N, \beta) = \sum_l \exp(-\beta E_l), \tag{70}$$

where the summation index l is restricted to the symmetry-allowed eigenstates $\{|\psi_l\rangle\}$ of $\hat{\mathcal{H}}$.

The Gibbs entropy of the quantum canonical ensemble can be written

$$S = k_B \beta \langle \hat{\mathcal{H}} \rangle + k_B \ln Z(V, N, \beta), \tag{71}$$

where

$$\langle \hat{\mathcal{H}} \rangle = \mathrm{Tr}\, \hat{\rho} \hat{\mathcal{H}} \tag{72a}$$

$$= \sum_m P(\psi_m) E_m \tag{72b}$$

is the average energy of the system.

The result given by Eq. (71) for the Gibbs entropy is identical in form to the result for the Gibbs entropy of a classical canonical ensemble, except that the average energy and the canonical partition function are to be computed using quantum mechanics rather than classical mechanics. As with the canonical probability density ρ for classical systems, the canonical density operator $\hat{\rho}$ for quantum systems represents the distribution of maximum Gibbs entropy consistent with the given average energy $\langle \hat{\mathcal{H}} \rangle$.

In view of the formal identity of the expressions for the Gibbs entropy of quantum and classical canonical ensembles, all the previously made formal connections between the thermodynamics and the statistical mechanics of closed, isothermal classical systems also apply to closed, isothermal quantum systems. [See Eqs. (60)–(66b).] One

need only replace the classical ensemble averages by quantum ensemble averages and reinterpret the classical canonical partition function as a quantum canonical partition function.

C. Isobaric–Isothermal Ensemble

1. Classical Systems

Consider a system made up of a fixed number of particles N enclosed in a motionless container with a variable volume V. The pressure P and temperature T of the system are held fixed. The macroscopic state of this system is called an isobaric–isothermal ensemble.

The probability density ρ for a classical isobaric–isothermal ensemble is given by

$$\rho(\vec{p}^N, \vec{q}^N; V) = \Delta^{-1}(P, N, \beta) \exp\{-\beta[\mathcal{H}(\vec{p}^N, \vec{q}^N; V) + PV]\}, \quad (73)$$

where β and P are parameters to be determined and

$$\Delta(P, N, \beta) = \int dV \int d^{6N}\Gamma \times \exp\{-\beta[\mathcal{H}(\vec{p}^N, \vec{q}^N; V) + PV]\} \quad (74)$$

is the isobaric–isothermal partition function. The probability density $\rho(\vec{p}^N, \vec{q}^N; V)$ is normalized as follows:

$$\int dV \int d^{6N}\Gamma \, \rho(\vec{p}^N, \vec{q}^N; V) = 1. \quad (75)$$

In writing the above equations, we have explicitly indicated the volume dependence of the probability density $\rho(\vec{p}^N, \vec{q}^N; V)$ and classical Hamiltonian $\mathcal{H}(\vec{p}^N, \vec{q}^N; V)$ to reflect the fact that the system volume V is variable.

The previously given forms for the ensemble average of classical dynamical variables and the Gibbs entropy associated with classical ensembles must be modified to accommodate the isobaric–isothermal ensemble. More specifically, we write

$$\langle O \rangle = \int dV \int d^{6N}\Gamma \, \rho(\vec{p}^N, \vec{q}^N; V) O(\vec{p}^N, \vec{q}^N; V) \quad (76)$$

and

$$S = -k_B \int dV \int d^{6N}\Gamma \, \rho(\vec{p}^N, \vec{q}^N; V) \ln \rho(\vec{p}^N, \vec{q}^N; V). \quad (77)$$

The Gibbs entropy of the isobaric–isothermal distribution given by Eq. (73) can be written

$$S = k_B \beta \langle \mathcal{H} \rangle + k_B \beta P \langle V \rangle + k_B \ln \Delta(P, N, \beta), \quad (78)$$

where $\langle \mathcal{H} \rangle$ and $\langle V \rangle$, respectively, represent the average energy and average volume of the system. The isobaric–isothermal distribution is the distribution of maximum Gibbs entropy consistent with the given average energy $\langle \mathcal{H} \rangle$ and average volume $\langle V \rangle$.

Equation (78) can be rearranged to read

$$\langle \mathcal{H} \rangle + P \langle V \rangle = -\beta^{-1} \ln \Delta(P, N, \beta) + (k_B \beta)^{-1} S. \quad (79)$$

This result is similar to the thermodynamic relations

$$H = E + PV \quad (80a)$$

$$= G + TS, \quad (80b)$$

where H is the enthalpy, E the internal energy, P the pressure, V the volume, G the Gibbs free energy, T the temperature, and S the entropy.

Assuming the Gibbs entropy of the isobaric–isothermal ensemble is equivalent to the thermodynamic entropy, the average energy $\langle \mathcal{H} \rangle$ and average volume $\langle V \rangle$ are equivalent to their thermodynamic analogues, $\beta = 1/k_B T$, and the macroscopic parameter P is equivalent to the pressure, we can write

$$G = -\beta^{-1} \ln \Delta(P, N, \beta). \quad (81)$$

This equation connects the Gibbs free energy G to microscopic interactions through the isobaric–isothermal partition function $\Delta(P, N, \beta)$.

The Gibbs free energy G is the thermodynamic characteristic function for the variables N, P, and T:

$$dG = -S\,dT + V\,dP + \sum_l \mu_l dN_l, \quad (82)$$

where

$$S = \left(-\frac{\partial G}{\partial T}\right)_{P,N}, \quad (83)$$

$$V = \left(\frac{\partial G}{\partial P}\right)_{T,N}, \quad (84)$$

and

$$\mu_l = \left(\frac{\partial G}{\partial N_l}\right)_{T,P,N'}. \quad (85)$$

In the above, S is the entropy, T the temperature, V the volume, P the pressure, μ_l the chemical potential of component l, and N_l the number of particles of component l.

Making use of the above results, one can construct the following statistical mechanical relations for the thermodynamic quantities S, V, and μ_l:

$$S = k_B T[\partial \ln \Delta(P, N, \beta)/\partial T]_{P,N} + k_B \ln \Delta(P, N, \beta), \quad (86)$$

$$V = -k_B T[\partial \ln \Delta(P, N, \beta)/\partial P]_{T,N}, \quad (87)$$

and

$$\mu_l = -k_B T[\partial \ln \Delta(P, N, \beta)/\partial N_l]_{T,P,N'}. \quad (88)$$

An expression connecting the enthalpy H to microscopic interactions can be obtained by differentiating $\ln \Delta(P, N, \beta)$ with respect to β:

$$H = -[\partial \ln \Delta(P, N, \beta)/\partial\beta]_{P,N}. \tag{89}$$

Differentiating $\ln \Delta(P, N, \beta)$ a second time with respect to β gives us the following expressions for the fluctuations in the variable $(\mathcal{H} + PV)$:

$$\begin{aligned}
&\langle[(\mathcal{H} + PV) - \langle(\mathcal{H} + PV)\rangle]^2\rangle \\
&\quad = \langle(\mathcal{H} + PV)^2\rangle - \langle(\mathcal{H} + PV)\rangle^2 && \text{(90a)} \\
&\quad = [\partial^2 \ln \Delta(P, N, B/\partial\beta^2]_{P,N} && \text{(90b)} \\
&\quad = -[\partial\langle(\mathcal{H} + PV)\rangle/\partial\beta]_{P,N}. && \text{(90c)}
\end{aligned}$$

Since the enthalpy H is equivalent to the ensemble average $\langle(\mathcal{H} + PV)\rangle$, these results enable us to connect the heat capacity $C_P = (\partial H/\partial T)_P$ at constant pressure to the fluctuations in the variable $(\mathcal{H} + PV)$ and microscopic interactions through the relations

$$\begin{aligned}
C_P &= k_B\beta^2\langle[(\mathcal{H} + PV) - \langle(\mathcal{H} + PV)\rangle]^2\rangle && \text{(91a)} \\
&= k_B\beta^2[\partial^2 \ln \Delta(P, N, \beta)/\partial\beta^2]_{P,N}. && \text{(91b)}
\end{aligned}$$

One might notice that the dimensions in Eq. (77) are incorrect. This problem can be corrected by introducing the scaled quantities $v = V/V_s$, $\rho_v(\vec{p}^N, \vec{q}^N) = V_s\rho(\vec{p}^N, \vec{q}^N; V)$, and $\Delta_v(P, N, \beta) = \Delta(P, N, \beta)/V_s$, where V_s is a suitable volume scale. If one reinterprets $\rho(\vec{p}^N, \vec{q}^N; V)$ as $\rho_v(\vec{p}^N, \vec{q}^N)$ and V as v in Eq. (77), this equation will be dimensionally correct. The results given by Eqs. (78), (79), (81), (86)–(89), (90b), and (91b) will also be dimensionally correct when $\Delta(P, N, \beta)$ is reinterpreted as $\Delta_v(P, N, \beta)$.

2. Quantum Systems

As for classical systems, the isobaric–isothermal ensemble is used to characterize the macroscopic state of a closed, isothermal quantum system with a variable volume and fixed pressure. The semiclassical (part classical, part quantum) density operator $\hat{\rho}$ for a quantum isobaric–isothermal ensemble is given by

$$\hat{\rho}(V) = \Delta^{-1}(P, N, \beta)\exp\{-\beta[\hat{\mathcal{H}}(V) + PV]\}, \tag{92}$$

where P is the pressure, $N = \{N_j\}$ the collection of particle numbers, $\beta = 1/k_BT$, $\hat{\mathcal{H}}(V)$ the Hamiltonian, V the volume (a classical variable), and

$$\Delta(P, N, B) = \int dV\,\mathrm{Tr}\exp\{-\beta[\hat{\mathcal{H}}(V) + PV]\} \tag{93}$$

the isobaric–isothermal partition function. The semiclassical density operator $\hat{\rho}(V)$ is normalized as follows:

$$\int dV\,\mathrm{Tr}\,\hat{\rho}(V) = 1. \tag{94}$$

The previously given forms for the ensemble average of a quantum dynamical variable and the Gibbs entropy associated with quantum ensembles must be modified to accommodate the isobaric–isothermal ensemble. More specifically, we write

$$\begin{aligned}
\langle\hat{O}\rangle &= \int dV \sum_m \rho\big(\psi_m^V\big)O\big(\psi_m^V\big) && \text{(95a)} \\
&= \int dV\,\mathrm{Tr}\,\hat{\rho}(V)\hat{O}(V) && \text{(95b)}
\end{aligned}$$

and

$$\begin{aligned}
S &= -k_B\int dV\,\mathrm{Tr}\,\hat{\rho}(V)\ln\hat{\rho}(V) && \text{(96a)} \\
&= -k_B\int dV \sum_m \rho\big(\psi_m^V\big)\ln\rho\big(\psi_m^V\big), && \text{(96b)}
\end{aligned}$$

where

$$O\big(\psi_m^V\big) = \big\langle\psi_m^V\big|\hat{O}(V)\big|\psi_m^V\big\rangle \tag{97}$$

and

$$\begin{aligned}
\rho\big(\psi_m^V\big) &= \big\langle\psi_m^V\big|\hat{\rho}(V)\big|\psi_m^V\big\rangle && \text{(98a)} \\
&= \Delta^{-1}(P, N, \beta)\exp\big[-\beta\big(E_m^V + PV\big)\big], && \text{(98b)}
\end{aligned}$$

with

$$\Delta(P, N, \beta) = \int dV \sum_l \exp\big[-\beta\big(E_l^V + PV\big)\big]. \tag{99}$$

It should be noted that both the eigenstates $\{|\psi_m^V\rangle\}$ and eigenvalues $\{E_m^V\}$ of the Hamiltonian $\mathcal{H}(V)$ are parametrically dependent on the system volume V.

The Gibbs entropy of the quantum isobaric–isothermal ensemble can be written

$$S = k_B\beta\langle\hat{\mathcal{H}}\rangle + k_B\beta P\langle V\rangle + k_B\ln\Delta(P, N, \beta), \tag{100}$$

where $\langle\hat{\mathcal{H}}\rangle$ and $\langle V\rangle$, respectively, are the average energy and average volume. This result is identical in form to the Gibbs entropy for a classical isobaric–isothermal ensemble, except the ensemble averages and partition function are to be computed using quantum mechanics rather than classical mechanics. As with the isobaric–isothermal probability density ρ for classical systems, the

isobaric–isothermal semiclassical density operator $\hat{\rho}(V)$ for quantum systems represents the distribution of maximum Gibbs entropy consistent with the given average energy $\langle\hat{\mathcal{H}}\rangle$ and average volume $\langle V\rangle$.

In view of the formal identity of the expressions for the Gibbs entropy of quantum and classical isobaric–isothermal ensembles, all the previously made formal connections between the thermodynamics and the statistical mechanics of closed, isothermal classical systems with variable volume and fixed pressure also apply to closed, isothermal quantum systems with variable volume and fixed pressure. [See Eqs. (86)–(91b).] One need only replace the classical ensemble averages by quantum ensemble averages and reinterpret the classical isobaric–isothermal partition function as a quantum isobaric–isothermal partition function.

It should be noted that Eqs. (96a) and (96b) suffer from the same dimensionality problem as Eq. (77) for the case of classical systems. Similar to the classical case, this problem can be corrected by introducing the scaled quantities $v = V/V_s$, $\hat{\rho}_v = V_s\,\hat{\rho}(V)$, and $\Delta_v(P, N, \beta) = \Delta(P, N, \beta)/V_s$, where V_s is a suitable volume scale. If one reinterprets $\hat{\rho}(V)$ as $\hat{\rho}_v$ and V as v in Eqs. (96a) and (96b), these equations will be dimensionally correct. The result given by Eq. (100) will also be dimensionally correct when $\Delta(P, N, \beta)$ is reinterpreted as $\Delta_v(P, N, \beta)$.

D. Grand Canonical Ensemble

1. Classical Systems

Thus far, our discussion of ensembles has been limited to closed systems, i.e., systems for which the number of particles N is fixed. Now we consider an open system, i.e., a system that can exchange particles with its surroundings. The volume V of the system is fixed and the temperature is maintained at the temperature T by keeping the system in thermal contact with a large heat bath (at the temperature T) with which it is able to exchange both energy and particles. Such a system is called an open, isothermal system. Its macroscopic state is called a grand canonical ensemble.

The probability density ρ for a classical grand canonical ensemble is given by

$$\rho(\vec{p}^N, \vec{q}^N) = \Xi^{-1}(V, \beta, \mu)\exp\left\{-\beta\left[\mathcal{H}(\vec{p}^N, \vec{q}^N) - \sum_l \mu_l N_l\right]\right\}, \tag{101}$$

where β and $\mu = \{\mu_l\}$ are parameters to be determined and

$$\Xi(V, \beta, \mu) = \sum_N \int d^{6N}\Gamma \exp\left\{-\beta\left[\mathcal{H}(\vec{p}^N, \vec{q}^N) - \sum_l \mu_l N_l\right]\right\} \tag{102}$$

is the grand canonical partition function. The probability density $\rho(\vec{p}^N, \vec{q}^N)$ is normalized as follows:

$$\sum_N \int d^{6N}\Gamma\, \rho(\vec{p}^N, \vec{q}^N) = 1. \tag{103}$$

In the above, the summation over N is intended to represent a summation over all of the particle numbers $N = \{N_j\}$. We have used a single summation index N to simplify the notation. Hereafter, this notation should be understood.

The previously given forms for the ensemble average of a classical dynamical variable and the Gibbs entropy associated with classical ensembles must be modified to accommodate the grand canonical ensemble. More specifically, we write

$$\langle O\rangle = \sum_N \int d^{6N}\Gamma\, \rho(\vec{p}^N, \vec{q}^N) O(\vec{p}^N, \vec{q}^N) \tag{104}$$

and

$$S = -k_B \sum_N \int d^{6N}\Gamma\, \rho(\vec{p}^N, \vec{q}^N) \ln\, \rho(\vec{p}^N, \vec{q}^N). \tag{105}$$

The Gibbs entropy of the classical grand canonical ensemble can be written

$$S = k_B\beta\langle\mathcal{H}\rangle - k_B\beta \sum_l \mu_l \langle N_l\rangle + k_B \ln \Xi(V, \beta, \mu), \tag{106}$$

where $\langle\mathcal{H}\rangle$ denotes the average energy and $\langle N_l\rangle$ the average number of particles of component l. The grand canonical distribution is the distribution of maximum Gibbs entropy consistent with the given average energy $\langle\mathcal{H}\rangle$ and average particle numbers $\{\langle N_l\rangle\}$.

Equation (106) can be rearranged to read

$$\langle\mathcal{H}\rangle + \beta^{-1} \ln \Xi(V, \beta, \mu) = (k_B\beta)^{-1} S + \sum_l \mu_l \langle N_l\rangle. \tag{107}$$

This result resembles the thermodynamic relation

$$E + PV = TS + \sum_l \mu_l N_l, \tag{108}$$

where E is the internal energy, P the pressure, V the volume, T the temperature, S the thermodynamic entropy, μ_l the chemical potential of component l, and N_l the number of particles of component l.

Assuming the Gibbs entropy of the grand canonical ensemble is equivalent to the thermodynamic entropy, the average energy $\langle \mathcal{H} \rangle$ and average particle numbers $\{\langle N_l \rangle\}$ are equivalent to their thermodynamic analogues, $\beta = 1/k_B T$, and the macroscopic parameter μ_l in Eq. (107) is equal to the chemical potential of component l, we can write

$$PV = \beta^{-1} \ln \Xi(V, \beta, \mu). \tag{109}$$

This result connects the product PV to microscopic interactions through the grand canonical partition function $\Xi(V, \beta, \mu)$.

The product PV is the thermodynamic characteristic function for the variables V, T, and μ_l:

$$d(PV) = S\,dT + \sum_l N_l\,d\mu_l + P\,dV, \tag{110}$$

where

$$S = [\partial(PV)/\partial T]_{\mu,V}, \tag{111}$$

$$N_l = [\partial(PV)/\partial \mu_l]_{V,T,\mu'}, \tag{112}$$

and

$$P = [\partial(PV)/\partial V]_{T,\mu}. \tag{113}$$

In the above, P is the pressure, V the volume, S the entropy, T the temperature, N_l the number of particles of component l, and μ_l the chemical potential of component l. The symbol μ' is used to indicate that the chemical potentials $\mu' = \{\mu_j; j \neq l\}$ are fixed. Hereafter, this notation should be understood.

Making use of the above results, one can construct the following statistical mechanical relations for the thermodynamic quantities S, N_l, and P:

$$S = k_B T[\partial \ln \Xi(V, \beta, \mu)/\partial T]_{V,\mu} + k_B \ln \Xi(V, \beta, \mu), \tag{114}$$

$$N_l = k_B T[\partial \ln \Xi(V, \beta, \mu)/\partial \mu_l]_{V,T,\mu'}, \tag{115}$$

and

$$P = k_B T[\partial \ln \Xi(V, \beta, \mu)/\partial V]_{\mu,T} \tag{116a}$$

$$= (k_B T/V) \ln \Xi(V, \beta, \mu). \tag{116b}$$

An expression for the average number of particles $\langle N_l \rangle$ of component l can be obtained by differentiating $\ln \Xi(V, \beta, \mu)$ with respect to μ_l:

$$\langle N_l \rangle = \beta^{-1} \partial \ln \Xi(V, \beta, \mu)/\partial \mu_l. \tag{117}$$

Differentiating a second time with respect to μ_l yields the following expressions for the particle number fluctutations in the system:

$$\langle [N_l - \langle N_l \rangle]^2 \rangle$$

$$= \langle N_l^2 \rangle - \langle N_l \rangle^2 \tag{118a}$$

$$= \beta^{-2} \left[\partial^2 \ln \Xi(V, \beta, \mu)/\partial \mu_l^2\right]_{V,T,\mu'} \tag{118b}$$

$$= \beta^{-1} (\partial \langle N_l \rangle/\partial \mu_l)_{V,T,\mu'}. \tag{118c}$$

For a single-component system, we can use these results and the thermodynamic relation $N(\partial \mu/\partial N)_{V,T} = V(\partial P/\partial N)_{V,T}$ to connect the isothermal compressibility $\kappa_T = -(1/V)(\partial V/\partial P)_{N,T}$ to particle number fluctuations and microscopic interactions:

$$\kappa_T = \frac{V}{k_B T} \frac{\langle [N - \langle N \rangle]^2 \rangle}{\langle N \rangle^2} \tag{119a}$$

$$= \frac{V}{k_B T} \frac{\partial^2 \ln \Xi(V, \beta, \mu)/\partial \mu^2}{[\partial \ln \Xi(V, \beta, \mu)/\partial \mu]^2}. \tag{119b}$$

2. Quantum Systems

As for classical systems, the grand canonical ensemble is used to characterize the macroscopic state of an open, isothermal quantum system. The density operator $\hat{\rho}$ for a quantum grand canonical ensemble is given by

$$\hat{\rho}(\hat{N}) = \Xi^{-1}(V, \beta, \mu) \exp\left\{-\beta\left[\hat{\mathcal{H}}(\hat{N}) - \sum_l \mu_l \hat{N}_l\right]\right\}, \tag{120}$$

where V is the volume, $\beta = 1/k_B T$, $\mu = \{\mu_l\}$ the collection of chemical potentials, $\hat{\mathcal{H}}(\hat{N})$ the Hamiltonian, $\hat{N} = \{\hat{N}_l\}$ the collection of particle number operators for the components of the system, and

$$\Xi(V, \beta, \mu) = \mathrm{Tr}\, \exp\left\{-\beta\left[\hat{\mathcal{H}}(\hat{N}) - \sum_l \mu_l \hat{N}_l\right]\right\} \tag{121}$$

the grand canonical partition function.

The eigenstates $\{|\psi_k^N\rangle\}$ of the system Hamiltonian $\hat{\mathcal{H}}(\hat{N})$ will also be eigenstates of the particle number operators $\hat{N}_l$ when the operators $\hat{\mathcal{H}}(\hat{N})$ and $\hat{N}_l$ commute. For such cases, we write

$$\hat{\mathcal{H}}(\hat{N})|\psi_k^N\rangle = E_k^N |\psi_k^N\rangle \tag{122}$$

and

$$\hat{N}_l |\psi_k^N\rangle = N_l |\psi_k^N\rangle. \tag{123}$$

In the above, $|\psi_k^N\rangle$ is an N-particle state vector satisfying the proper symmetry requirements with respect to particle permutation and E_k^N is the energy of state $|\psi_k^N\rangle$.

In general, the particle number operators $\hat{N}_l$ do not commute with the Hamiltonian $\hat{\mathcal{H}}(\hat{N})$ of a system. This will be the case, for example, in multicomponent reactive systems. For such cases, the Hamiltonian contains interaction

terms that give rise to the transformation of one set of particles into another set of particles. Hereafter, we shall not consider such complications and restrict our considerations to nonreactive systems.

The previously given forms for the ensemble average of a quantum dynamical variable and the Gibbs entropy associated with quantum ensembles must be modified to accommodate the grand canonical ensemble. More specifically, we write

$$\langle \hat{O} \rangle = \sum_N \mathrm{Tr}\, \hat{\rho}(N) \hat{O}(N) \tag{124a}$$

$$= \sum_N \sum_k P\left(\psi_k^N\right) O\left(\psi_k^N\right) \tag{124b}$$

and

$$S = -k_\mathrm{B} \sum_N \mathrm{Tr}\, \hat{\rho}(N) \ln \hat{\rho}(N) \tag{125a}$$

$$= -k_\mathrm{B} \sum_N \sum_k P\left(\psi_k^N\right) \ln P\left(\psi_k^N\right), \tag{125b}$$

where

$$O\left(\psi_k^N\right) = \left\langle \psi_k^N \middle| \hat{O}(\hat{N}) \middle| \psi_k^N \right\rangle. \tag{126}$$

In the above,

$$P\left(\psi_k^N\right) = \Xi^{-1}(V, \beta, \mu) \exp\left[-\beta \left(E_k^N - \sum_l \mu_l N_l \right) \right] \tag{127}$$

is the probability of finding the system in the N-particle state $|\psi_k^N\rangle$, where

$$\Xi(V, \beta, \mu) = \sum_N \sum_k \exp\left[-\beta \left(E_k^N - \sum_l \mu_l N_l \right) \right]. \tag{128}$$

The Gibbs entropy of the quantum grand canonical ensemble can be written

$$S = k_\mathrm{B} \beta \langle \hat{\mathcal{H}} \rangle - k_\mathrm{B} \beta \sum_l \mu_l \langle \hat{N}_l \rangle + k_\mathrm{B} \ln \Xi(V, \beta, \mu), \tag{129}$$

where $\langle \hat{\mathcal{H}} \rangle$ denotes the average energy and $\langle \hat{N}_l \rangle$ the average number of particles of component l. This result is identical in form to the Gibbs entropy for a classical grand canonical ensemble, except the ensemble averages and partition function are to be computed using quantum mechanics rather than classical mechanics. As with the grand canonical probability density ρ for classical systems, the grand canonical density operator $\hat{\rho}$ for quantum systems represents the distribution of maximum Gibbs entropy consistent with the given average energy $\langle \hat{\mathcal{H}} \rangle$ and average particle numbers $\{\langle \hat{N}_l \rangle\}$.

In view of the formal identity of the expressions for the Gibbs entropy of quantum and classical grand canonical ensembles, all of the previously made formal connections between the thermodynamics and the statistical mechanics of open, isothermal classical systems also apply to open, isothermal quantum systems. [See Eqs. (114)–(119b).] One need only replace the classical ensemble averages by quantum ensemble averages and reinterpret the classical grand canonical partition function as a quantum grand canonical partition function.

E. On the Equivalence of Gibbsian Ensembles

The Gibbsian ensembles discussed earlier were defined by specifying certain physical conditions. For example, the canonical ensemble is defined by keeping the system in thermal contact with a large heat bath at some temperature and fixing the particle numbers and volume. From a theoretical point of view, the results for the various Gibbsian ensembles hold only for the conditions under which the Gibbsian ensembles are defined. Nonetheless, in actual applications of statistical mechanics, investigators often choose a Gibbsian ensemble on the basis of mathematical convenience rather than on the basis of the physical conditions under which a system is found.

The replacing of one ensemble by another is often rationalized by appealing to scaling arguments that lead to the conclusion that the size of the fluctuations in the various ensembles is vanishingly small. Of course, there are cases for which the fluctuations may not be small, for example, when two phases coexist and when a critical point is realized. In order to rigorously establish the thermodynamic equivalence of the Gibbsian ensembles, one must demonstrate that the ensembles give the same results for all physically relevant quantities.

III. INFORMATION THEORY AND STATISTICAL MECHANICS

A. Isomorphism between Information Theory and Statistical Mechanics

Consider a system that can be described in terms of a set of events. We say that an event occurs when some variable(s) used to characterize the event assumes some value(s). Generally, we lack sufficient information to specify the exact probability distribution of events. Rather, we possess only a limited amount of information, such as the average value(s) of the variable(s) characterizing the events. Nonetheless, it is desirable to make use of our limited information to make educated guesses about the probability of an event and to make estimates of the properties

of our system. This is the basic problem confronting us in statistical mechanics.

The problem of determining the probability distribution of events requires us to find a least-bias distribution that agrees with the given information about a system. The resolution of this problem underwent a great advance with the advent of Shannon's information theory. Shannon showed that there exists a unique measure of the statistical uncertainty in a discrete probability distribution $\{P(j)\}$, which is in agreement with the intuitive notion that a broad distribution represents more uncertainty than a narrow distribution. This measure is called the missing information, which is defined by

$$I = -c \sum_j P(j) \ln P(j), \tag{130}$$

where c is a positive constant. The missing information I is nonnegative and additive for independent sources of uncertainty.

The additivity property of the missing information I can be illustrated by considering a distribution $\{P_{12}(j,k)\}$ of joint probabilities $P_{12}(j,k)$ of the form $P_{12}(j,k) = P_1(j)P_2(k)$, where $\{P_1(j)\}$ and $\{P_2(k)\}$ are uncorrelated probability distributions. The missing information

$$I = -c \sum_{j,k} P_{12}(j,k) \ln P_{12}(j,k) \tag{131}$$

for the distribution $\{P_{12}(j,k)\}$ can be expressed as

$$I = I_1 + I_2, \tag{132}$$

where

$$I_1 = -c \sum_j P_1(j) \ln P_1(j) \tag{133}$$

and

$$I_2 = -c \sum_k P_2(k) \ln P_2(k). \tag{134}$$

Equation (132) expresses the additivity property of the missing information for two independent sources of uncertainty.

The generalization of the expression given by Eq. (130) for the missing information I to include continuous distributions can be accomplished by treating the continuous problem as a limiting case of the discrete problem. For a continuous distribution of events, we write

$$I = -c \int dx\, \rho(x) \ln \rho(x), \tag{135}$$

where $\rho(x)$ is a normalized probability density. As for discrete distributions, the missing information is nonnegative and additive for independent sources of uncertainty.

In information theory, one chooses the probability distribution on the basis of the available information and requires the distribution to possess a maximum in the missing information. This is the only unbiased distribution that we can postulate. The use of any other distribution would be equivalent to making arbitrary assumptions about unavailable information. The actual construction of information-theoretic distributions is accomplished by treating the linearly independent pieces of known information as constraints and utilizing the method of Lagrange undetermined multipliers.

Information-theoretic distributions constructed in the above-described manner are identical in form to the distributions for the various quantum and classical Gibbsian ensembles discussed earlier. This can be traced to the formal equivalence of the missing information I and the Gibbs entropy S. As with information-theoretic distributions, the Gibbsian distributions are maximum entropy distributions consistent with the information defining the macroscopic state of a system.

The above-described isomorphism between information theory and statistical mechanics has suggested to a number of investigators that statistical mechanics should be reinterpreted in such a way that information theory can be used to justify its formal structure. This point of view was pioneered by Jaynes. Within the context of the information-theoretic approach to statistical mechanics, entropy and entropy maximization are fundamental concepts. The information-theoretic approach to statistical mechanics not only leads to significant conceptual and mathematical simplification of the subject, but also frees us from such hypotheses as ergodicity and equal a priori probabilities. In addition, it provides a clear path for the formal development of statistical mechanics and thermodynamics without requiring us to appeal to phenomenological equations in order to render the statistical mechanical formalism meaningful.

B. Maximum Entropy Principle

1. Classical Version

If we interpret Gibbs entropy in the same spirit as the missing information of information theory, it can be viewed as a measure of statistical uncertainty. Adopting this point of view, it seems natural to treat the following principle as a basic postulate of classical equilibrium statistical mechanics.

Maximum entropy principle (classical version). The probability density ρ used to describe the macroscopic state of a system must represent all available information and correspond to the maximum entropy S distribution consistent with this information, where

$$S = -k_B \operatorname{Tr} \rho \ln \rho. \tag{136}$$

In the above, we have used the symbol Tr to indicate a classical trace, i.e., a sum over all accessible classical states. This sum includes an integration over the accessible phase space, and any other integrations and/or summations over classical variables required to define the macroscopic state of the system.

The implementation of the maximum entropy principle can be illustrated by considering a system for which we possess information about the average values $\{\langle O_j\rangle;\ j = 0,\ldots,n\}$ of some set of observables, where

$$\langle O_j\rangle = \mathrm{Tr}\,\rho O_j. \tag{137}$$

The classical variable $O_0 = 1$. Thus, $\langle O_0\rangle = 1$ is a statement of the requirement that ρ be normalized to unity.

According to the maximum entropy principle, the probability density ρ describing the macroscopic state of our system must represent all available information, $\{\langle O_j\rangle;\ j = 0,\ldots,n\}$, and correspond to a maximum in Gibbs entropy S. The actual construction of ρ can be accomplished by adopting a procedure due to Lagrange that allows one to change a constrained extremum problem into an unconstrained extremum problem.

Adopting the method of Lagrange, we seek an unconstrained extremum of the auxilary function $\mathcal{L}$, called the Lagrangian, which is defined by

$$\mathcal{L} = S + k_{\mathrm{B}}(\Omega+1)\langle 1\rangle - k_{\mathrm{B}}\sum_{j=1}^{n}\Lambda_j\langle O_j\rangle, \tag{138}$$

where $-k_{\mathrm{B}}(\Omega+1)$ and $\{k_{\mathrm{B}}\Lambda_j;\ j = 1,\ldots,n\}$ are Lagrange multipliers.

The variation $\delta\mathcal{L}$ in the Lagrangian $\mathcal{L}$ due to the variation $\delta\rho$ in the probability density ρ is given by

$$\delta\mathcal{L} = -k_{\mathrm{B}}\,\mathrm{Tr}\,\delta\rho\left(\ln\rho - \Omega + \sum_{j=1}^{n}\Lambda_j O_j\right). \tag{139}$$

Since we seek an unconstrained extremum of $\mathcal{L}$, the variation $\delta\mathcal{L}$ in $\mathcal{L}$ must vanish for arbitrary variations $\delta\rho$ in ρ. It follows that

$$\rho = \exp\left(\Omega - \sum_{j=1}^{n}\Lambda_j O_j\right). \tag{140}$$

The probability density ρ is expressed in terms of the Lagrange parameters Ω and $\{\Lambda_j;\ j = 1,\ldots,n\}$. These quantities can be determined from the constraints, i.e., Eq. (137) for $j = 0,\ldots,n$.

Imposing the normalization constraint $\langle O_0\rangle = 1$, we obtain

$$\Omega = -\ln Z, \tag{141}$$

where

$$Z = \mathrm{Tr}\exp\left(-\sum_{j=1}^{n}\Lambda_j O_j\right). \tag{142}$$

The quantity Z can be regarded as a generalized partition function.

The remaining Lagrange parameters $\{\Lambda_j;\ j = 1,\ldots,n\}$ can be determined by solving the set of equations

$$\langle O_j\rangle = \mathrm{Tr}\exp\left(\Omega - \sum_{k=1}^{n}\Lambda_k O_k\right)O_j \tag{143}$$

or

$$\langle O_j\rangle = \partial\Omega/\partial\Lambda_j, \tag{144}$$

for $j = 1,\ldots,n$, with Ω expressed in terms of $\{\Lambda_j;\ j = 1,\ldots,n\}$ by Eqs. (141) and (142).

The equations for the Lagrange parameters can be cast in the form

$$-\frac{\partial F}{\partial\Lambda_j} = \langle O_j\rangle \tag{145}$$

for $j = 0,\ldots,n$, where $\langle O_0\rangle = 1,\ \Lambda_0 = -\Omega$, and

$$F = \mathrm{Tr}\exp\left(-\sum_{k=0}^{n}\Lambda_k O_k\right) \tag{146}$$

is an implicit function of $\{\Lambda_k; k = 0,\ldots,n\}$.

A necessary condition for Eq. (145) to be solvable for the Λ_j with arbitrary $\langle O_j\rangle$ and a sufficient condition for the solution to be unique is that the matrix with the elements

$$\frac{\partial^2 F}{\partial\Lambda_j\partial\Lambda_k} = \langle O_j O_k\rangle \tag{147}$$

be regular. This requires the pieces of information $\{\langle O_j\rangle;\ j = 0,\ldots,n\}$ to be linearly independent. Assuming that this is indeed the case, the probability density ρ is unique.

The entropy S of the distribution ρ can be written

$$S = -k_{\mathrm{B}}\Omega + k_{\mathrm{B}}\sum_{j=1}^{n}\Lambda_j\langle O_j\rangle. \tag{148}$$

One can readily verify that ρ does indeed represent the maximum entropy distribution consistent with the available information $\{\langle O_j\rangle;\ j = 0,\ldots,n\}$.

From the above discussion, it should be clear that all of the classical probability densities for the various Gibbsian ensembles discussed earlier can be constructed by starting with the maximum entropy principle and making use of the information defining the macroscopic state of the system.

2. Quantum Version

The quantum version of the maximum entropy principle can be stated as follows.

Maximum entropy principle (quantum version). The density operator $\hat{\rho}$ used to describe the macroscopic state of a system must represent all available information and correspond to the maximum entropy S density operator consistent with this information, where the entropy S is defined by

$$S = -k_B \operatorname{Tr} \hat{\rho} \ln \hat{\rho}. \tag{149}$$

If classical variables are also required to specify the macroscopic state of the system, the trace includes the relevant integrations and/or summations over the classical variables.

The implementation of the quantum version of the maximum entropy principle can be accomplished in a fashion similar to the implementation of the classical version of this principle. Of course, one must be careful to account for the possible noncommutativity of the relevant quantum operators.

For a quantum system, our information is the set of averages $\{\langle \hat{O}_j \rangle; j = 0, \ldots, n\}$, where the operators $\{\hat{O}_j\}$ are assumed to be Hermitian. Making use of this information and the method of Lagrange, we obtain

$$\hat{\rho} = \exp\left(\Omega \hat{I} - \sum_{j=1}^{n} \Lambda_j \hat{O}_j\right), \tag{150}$$

where

$$\Omega = -\ln Z, \tag{151}$$

with

$$Z = \operatorname{Tr} \exp\left(-\sum_{j=1}^{n} \Lambda_j \hat{O}_j\right). \tag{152}$$

Similar to the classical case, one can demonstrate that the density operator $\hat{\rho}$ is unique provided the pieces of information $\{\langle \hat{O}_j \rangle; j = 0, \ldots, n\}$ are linearly independent. Also, one can readily verify that $\hat{\rho}$ does indeed represent the maximum entropy density operator consistent with the information defining the macroscopic state of the system.

The entropy S of the macroscopic state described by $\hat{\rho}$ can be written

$$S = -k_B \Omega + k_B \sum_{j=1}^{n} \Lambda_j \langle \hat{O}_j \rangle. \tag{153}$$

As expected, this result is identical in form to the classical result.

Of course, we require $\hat{\rho}$ to be a Hermitian operator with nonnegative eigenvalues. Since the operators $\{\hat{O}_j\}$ are assumed to be Hermitian, the Hermiticity requirement of $\hat{\rho}$ will be satisfied provided the Lagrange parameters $\{\Lambda_j; j = 1, \ldots, n\}$ are real. Assuming this to be the case, we find that the eigenvalues

$$P(\psi) = \exp\left(\Omega - \sum_{j=1}^{n} \Lambda_j \langle \hat{O}_j \rangle_\psi\right) \tag{154}$$

of $\hat{\rho}$ are real and nonnegative.

All of the density operators for the various Gibbsian ensembles discussed earlier can be constructed by starting from the quantum version of the maximum entropy principle and making use of the information defining the macroscopic state of the system.

C. Statistical Mechanical Basis for Equilibrium Thermodynamics

In the last section, we used the maximum entropy principle to construct the density operator $\hat{\rho}$ characterizing the macroscopic state of some general quantum system for which the information $\{\langle \hat{O}_j \rangle; j = 0, \ldots, n\}$ is available. Let us turn our attention to the relationship between the statistical mechanics and thermodynamics of this system.

Introducing the quantity

$$\Phi = -k_B \Omega, \tag{155}$$

we can rewrite the expression given by Eq. (153) for the entropy S of the macroscopic state described by $\hat{\rho}$ as

$$S = \Phi + k_B \sum_{j=1}^{n} \Lambda_j \langle \hat{O}_j \rangle. \tag{156}$$

As with Ω, the quantity Φ is a function of the Lagrange parameters $\{\Lambda_j; j = 1, \ldots, n\}$.

Making use of Eqs. (151), (152), and (155), one can demonstrate that

$$\Phi = k_B \ln Z \tag{157}$$

and

$$\frac{\partial \Phi}{\partial \Lambda_j} = -k_B \langle \hat{O}_j \rangle. \tag{158}$$

These equations can be combined to give

$$\langle \hat{O}_j \rangle = -\frac{\partial \ln Z}{\partial \Lambda_j}. \tag{159}$$

Substituting Eqs. (157) and (159) into Eq. (156), we obtain

$$S = k_B \ln Z - k_B \sum_{j=1}^{n} \Lambda_j \frac{\partial \ln Z}{\partial \Lambda_j}. \tag{160}$$

The relations given by Eqs. (157), (159), and (160) connect the quantities Φ, $\langle \hat{O}_j \rangle$, and S to microscopic interactions through the generalized partition function Z.

In view of Eq. (156), the differential dS can be written

$$dS = k_B \sum_{j=1}^{n} \left\{ \left[\frac{1}{k_B} \frac{\partial \Phi}{\partial \Lambda_j} + \langle \hat{O}_j \rangle \right] d\Lambda_j + \Lambda_j d\langle \hat{O}_j \rangle \right\}. \tag{161}$$

Making use of Eq. (158), we obtain

$$dS = k_B \sum_{j=1}^{n} \Lambda_j d\langle \hat{O}_j \rangle. \tag{162}$$

This result reveals that the partial derivatives of the entropy S are given by

$$\frac{\partial S}{\partial \langle \hat{O}_j \rangle} = k_B \Lambda_j. \tag{163}$$

Since the quantity Φ depends solely on the Lagrange parameters $\{\Lambda_j; j = 1, \ldots, n\}$, we can write

$$d\Phi = \sum_{j=1}^{n} \frac{\partial \Phi}{\partial \Lambda_j} d\Lambda_j. \tag{164}$$

In view of Eq. (158), this equation can be written

$$d\Phi = -k_B \sum_{j=1}^{n} \langle \hat{O}_j \rangle \, d\Lambda_j. \tag{165}$$

The macroscopic state of the system can be defined in terms of either the Lagrange parameters $\{\Lambda_j; j = 1, \ldots, n\}$ or the averages $\{\langle \hat{O}_j \rangle; j = 1, \ldots, n\}$. Both sets of variables represent the same information. If the Lagrange parameters are known, the set of equations given by (163) can be solved for the averages. If the averages are known, the set of equations given by (158) can be solved for the Lagrange parameters.

Interpreting the macroscopic state of the system as its thermodynamic state, we can regard the quantity Φ as a generalized thermodynamic potential. The Lagrange parameters $\{\Lambda_j; j = 1, \ldots, n\}$ can be thought to represent thermodynamic parameters. We shall refer to the averages $\{\langle \hat{O}_j \rangle; j = 1, \ldots, n\}$ as thermodynamic coordinates. As indicated above, the thermodynamic state of the system can be defined in terms of either the thermodynamic parameters or the thermodynamic coordinates.

One can demonstrate that the entropy S and the thermodynamic potential Φ are connected through the generalized Gibbs–Helmholtz relations

$$S = \Phi - \sum_{j=1}^{n} \Lambda_j \frac{\partial \Phi}{\partial \Lambda_j} \tag{166}$$

and

$$\Phi = S - \sum_{j=1}^{n} \langle \hat{O}_j \rangle \frac{\partial S}{\partial \langle \hat{O}_j \rangle}. \tag{167}$$

These results reveal that the entropy S and thermodynamic potential Φ are Legendre transforms of each other. Moreover, we find from Eq. (167) that Φ is a generalized Massieu function. Such functions play an important role in thermodynamics.

Making use of Eqs. (158) and (163) and the equalities

$$\frac{\partial^2 S}{\partial \langle \hat{O}_j \rangle \partial \langle \hat{O}_k \rangle} = \frac{\partial^2 S}{\partial \langle \hat{O}_k \rangle \partial \langle \hat{O}_j \rangle} \tag{168}$$

and

$$\frac{\partial^2 \Phi}{\partial \Lambda_j \partial \Lambda_k} = \frac{\partial^2 \Phi}{\partial \Lambda_k \partial \Lambda_j}, \tag{169}$$

we obtain the generalized Maxwell relations

$$\frac{\partial \Lambda_k}{\partial \langle \hat{O}_j \rangle} = \frac{\partial \Lambda_j}{\partial \langle \hat{O}_k \rangle} \tag{170}$$

and

$$\frac{\partial \langle \hat{O}_k \rangle}{\partial \Lambda_j} = \frac{\partial \langle \hat{O}_j \rangle}{\partial \Lambda_k}. \tag{171}$$

In thermodynamics, the Maxwell relations are used to connect the partial derivatives of a great diversity of thermodynamic quantities.

In general, the thermodynamic state of the system depends not only on the averages $\{\langle \hat{O}_j \rangle; j = 1, \ldots, n\}$, but also on external parameters $\{a_l; l = 1, \ldots, m\}$ that we have regarded as fixed. Such parameters may include volume, intensity of an external electric or magnetic field, etc.

In thermodynamics, external parameters are assumed to be adjustable quantities that can be varied in such a way that the system can pass through an ordered sequence of equilibrium states. A process of this type is called a quasi-static or reversible process. For an infinitesimal quasi-static process, dS vanishes.

In order to provide a more complete thermodynamic description of our system, we must consider quasi-static processes. It will be assumed that such processes exist, without any theoretical justification.

Assuming the external parameters are indeed adjustable quantities, the thermodynamic potential $\Omega = -\Phi/k_B$ can be regarded as a function of both the averages $\{\langle \hat{O}_j \rangle; j = 1, \ldots, n\}$ and the external parameters $\{a_l; l = 1, \ldots, m\}$. Then the differential $d\Omega$ is given by

$$d\Omega = \sum_{j=1}^{n} \frac{\partial \Omega}{\partial \Lambda_j} d\Lambda_j + \sum_{l=1}^{m} \frac{\partial \Omega}{\partial a_l} da_l. \tag{172}$$

Introducing the definition

$$f_l = -\frac{\partial \Omega}{\partial a_l} \tag{173}$$

and making use of the relation

$$\frac{\partial \Omega}{\partial \Lambda_j} = \langle \hat{O}_j \rangle, \tag{174}$$

we obtain

$$d\Omega = \sum_{j=1}^{n} \langle \hat{O}_j \rangle d\Lambda_j - \sum_{l=1}^{m} f_l da_l. \tag{175}$$

The quantity f_l can be thought to represent a generalized force arising from the variation in the generalized coordinate a_l.

It follows from Eq. (153) that the differential dS can be written

$$dS = -k_B d\Omega + k_B \sum_{j=1}^{n} [\Lambda_j d\langle \hat{O}_j \rangle + \langle \hat{O}_j \rangle d\Lambda_j]. \tag{176}$$

Substitution of Eq. (175) yields

$$dS = k_B \sum_{j=1}^{n} \Lambda_j d\langle \hat{O}_j \rangle + k_B \sum_{l=1}^{m} f_l da_l. \tag{177}$$

Assuming the system has undergone an infinitesimal quasi-static process, dS must vanish. Imposing this condition on Eq. (177), we obtain

$$\sum_{j=1}^{n} \Lambda_j d\langle \hat{O}_j \rangle = -\sum_{l=1}^{m} f_l da_l. \tag{178}$$

This result tells us how the averages $\{\langle \hat{O}_j \rangle; j = 1, \ldots, n\}$ adiabatically follow the external parameters $\{a_l; l = 1, \ldots, m\}$ in a quasi-static process.

If the average energy $\langle \hat{\mathcal{H}} \rangle$ of the system is included among the averages, $\langle \hat{O}_1 \rangle = \langle \hat{\mathcal{H}} \rangle$, we can recast Eq. (178) in the form

$$dE = -dW, \tag{179}$$

where $dE = d\langle \hat{\mathcal{H}} \rangle$ is the change in the average energy and

$$dW = \Lambda_1^{-1} \left[\sum_{l=1}^{m} f_l da_l + \sum_{j=2}^{n} \Lambda_j d\langle \hat{O}_j \rangle \right] \tag{180}$$

is the work done (energy expended) in changing the external parameters and the averages other than the average energy. Since the energy in a quasi-static process is recoverable by simply changing the averages and external parameters back to their original values, dW is often called reversible work.

In order for us to introduce the concepts of heat and heat transfer, we must include among the averages $\{\langle \hat{O}_j \rangle; j = 1, \ldots, n\}$ the average energy $\langle \hat{\mathcal{H}} \rangle$ and bring the system in thermodynamic contact with another system. Thus, we must consider a composite system comprised of two subsystems. The macroscopic state of the composite system is defined in terms of the external parameters $\{a_l^{(\alpha)}; l = 1, \ldots, m^{(\alpha)}\}$ and the averages $\{\langle \hat{O}_j^{(\alpha)} \rangle; j = 1, \ldots, n^{(\alpha)}\}$ for each of the subsystems, labeled as $\alpha = 1$ and 2. The average $\langle \hat{O}_1^{(\alpha)} \rangle$ is taken to represent the average energy $\langle \hat{\mathcal{H}}^{(\alpha)} \rangle$ of subsystem α.

For the sake of simplicity, we assume that the sets of variables $\{\langle \hat{O}_j^{(\alpha)} \rangle; j = 1, \ldots, n^{(\alpha)}\}$ and $\{a_l^{(\alpha)}; l = 1, \ldots, m^{(\alpha)}\}$ are identical for both subsystems and subject to the conservation relations

$$\langle \hat{O}_j \rangle = \langle \hat{O}_j^{(1)} \rangle + \langle \hat{O}_j^{(2)} \rangle \tag{181}$$

and

$$a_l = a_l^{(1)} + a_l^{(2)}, \tag{182}$$

where $\langle \hat{O}_j \rangle$ and a_l are fixed.

Before the two subsystems are brought into thermodynamic contact, they are statistically independent. Consequently, the values of the variables defining the macroscopic state of one subsystem are uncorrelated with the values of the variables defining the macroscopic state of the other subsystem. Thus, the entropy S of the composite system before thermodynamic contact can be written

$$S = S^{(1)} + S^{(2)}, \tag{183}$$

where

$$S^{(\alpha)} = -k_B \Omega^{(\alpha)} + k_B \Lambda_1^{(\alpha)} \langle \hat{\mathcal{H}}^{(\alpha)} \rangle + k_B \sum_{j=2}^{n^{(\alpha)}} \Lambda_j^{(\alpha)} \langle \hat{O}_j^{(\alpha)} \rangle \tag{184}$$

is the entropy $S^{(\alpha)}$ of subsystem α. The thermodynamic potential $\Omega^{(\alpha)}$ is given by

$$\Omega^{(\alpha)} = -\ln Z^{(\alpha)}, \tag{185}$$

where

$$Z^{(\alpha)} = \mathrm{Tr} \exp\left(-\Lambda_1^{(\alpha)} \hat{\mathcal{H}}^{(\alpha)} - \sum_{j=2}^{n^{(\alpha)}} \Lambda_j^{(\alpha)} \hat{O}_j^{(\alpha)} \right). \tag{186}$$

Now we bring the two subsystems into thermodynamic contact so that they can communicate via infinitesimal transfer processes described by the differentials $da_l^{(\alpha)}$ and $d\langle \hat{O}_j^{(\alpha)} \rangle$. The entropy change dS accompanying these infinitesimal transfer processes can be written

$$dS = dS^{(1)} + dS^{(2)}, \tag{187}$$

where the entropy change $dS^{(\alpha)}$ associated with subsystem α is given by

$$dS^{(\alpha)} = k_B \Lambda_1^{(\alpha)} \left[dE^{(\alpha)} + dW^{(\alpha)} \right]. \tag{188}$$

In the above, $dE^{(\alpha)} = d\langle \hat{\mathcal{H}}^{(\alpha)} \rangle$ is the change in the average energy of subsystem α and

$$dW^{(\alpha)} = \Lambda_1^{(\alpha)^{-1}} \left[\sum_{l=1}^{m^{(\alpha)}} f_l^{(\alpha)} da_l^{(\alpha)} + \sum_{j=2}^{n^{(\alpha)}} \Lambda_j^{(\alpha)} d\langle \hat{O}_j^{(\alpha)} \rangle \right] \tag{189}$$

is the work done by subsystem α in changing its external parameters and averages.

Making use of the conservation relations given by Eqs. (181) and (182), we recast Eq. (187) in the following equivalent form:

$$dS = k_B\left[\Lambda_1^{(1)} - \Lambda_1^{(2)}\right] d\langle\hat{\mathcal{H}}^{(1)}\rangle + k_B \sum_{j=2}^{n^{(1)}} \left[\Lambda_j^{(1)} - \Lambda_j^{(2)}\right] d\langle\hat{O}_j^{(1)}\rangle + k_B \sum_{l=1}^{m^{(1)}} \left[f_l^{(1)} - f_l^{(2)}\right] da_l^{(1)}. \quad (190)$$

If the composite system is in a state of equilibrium, the entropy change dS must vanish. Assuming this to be the case, we obtain

$$\Lambda_j^{(1)} = \Lambda_j^{(2)} \quad (191)$$

and

$$f_l^{(1)} = f_l^{(2)}. \quad (192)$$

These relations represent a set of conditions for equilibrium between subsystems 1 and 2.

It is quite possible for the composite system to be in a state of equilibrium when some or none of the thermodynamic parameters and external forces satisfy Eqs. (191) and (192). Equilibrium states of this type are realized when the imposed restraints prevent the occurrence of some or all of the transfer processes described by the differentials $d\langle\hat{O}_j^{(\alpha)}\rangle$ and $da_l^{(\alpha)}$. (If a given transfer process is forbidden, the differential representing that transfer process must be set equal to zero.) In view of the possibility of such restrictions, we conclude that equilibrium between the two subsystems requires identical thermodynamic parameters and identical external forces when the transfer processes associated with these quantities are allowed by the restraints placed on the composite system. This is a general thermodynamic criterion for equilibrium between the two systems.

If the two subsystems are not in equilibrium, we say that they are out of equilibrium or in a state of disequilibrium. When the composite system is out of equilibrium the second law of thermodynamics requires the entropy change dS associated with an infinitesimal process to be positive. After the composite system, originally in a state of disequilibrium, has exhausted all allowed spontaneous transfer processes, the entropy will reach a maximum and the conditions of equilibrium consistent with the imposed restraints will be realized. Any further changes in the composite system will be reversible and the equality $dS = 0$ will hold.

Assuming that the composite system has achieved a state of equilibrium for which $\Lambda_1^{(1)} = \Lambda_1^{(2)} = \Lambda_1$ (energy transfer is allowed), we conclude from Eqs. (187) and (188) and the condition $dS = 0$ that the following relation must hold for the composite system undergoing an infinitesimal reversible process:

$$\left[dE^{(1)} + dW^{(1)}\right] = -\left[dE^{(2)} + dW^{(2)}\right]. \quad (193)$$

If the energy increase $dE^{(1)}$ of subsystem 1 is not equal to the work done $dW^{(1)}$ (energy expended) by this subsystem, there is an energy mismatch given by

$$dQ^{(1)} = dE^{(1)} + dW^{(1)}. \quad (194)$$

This energy mismatch can be accounted for by an energy transfer $dQ^{(1)}$ from subsystem 2 to subsystem 1. The energy transfer $dQ^{(1)}$ is commonly called heat transfer.

Rearrangement of Eq. (194) gives the first law of thermodynamics:

$$dE^{(1)} = dQ^{(1)} - dW^{(1)}. \quad (195)$$

This law tells us that the energy stored in a system during an infinitesimal quasi-static process is the difference between the heat flux into the system and the work performed by the system.

In view of Eqs. (188) and (194), we can write the entropy change $dS^{(1)}$ of subsystem 1 as

$$dS^{(1)} = k_B \Lambda_1^{(1)} dQ^{(1)}, \quad (196)$$

where $k_B\Lambda_1^{(1)}$ plays the role of an integrating factor for converting the inexact differential $dQ^{(1)}$ into an exact differential $dS^{(1)}$. Defining the inverse of the integrating factor $k_B\Lambda_1^{(1)}$ as the absolute temperature $T^{(1)}$, we obtain for Eq. (196) the form

$$dS^{(1)} = dQ^{(1)}/T^{(1)}. \quad (197)$$

As in thermodynamics, we find that a quasi-static heat flux into a system can be associated with an increase in the entropy of that system.

The formal development given above constitutes a complete thermodynamic description of a general quantum system. Adopting a similar development for classical systems, one finds that the final results are identical in form to the quantum results. The only difference between the quantum and classical results lies at the microscopic level in the statistical mechanical expressions for the thermodynamic quantities.

With the maximum entropy principle and the notion of quasi-static processes at our disposal, we showed that the thermodynamics of a system emerges from the statistical mechanics of that system. Unlike conventional approaches to equilibrium statistical mechanics, the information-theoretic approach does not require us to appeal to the phenomenological equations of thermodynamics in order to render the statistical mechanical formalism meaningful.

The application of the above-described results to the development of the thermodynamics of the various systems discussed in Section 2 is straightforward.

IV. LIOUVILLE DESCRIPTION OF TIME EVOLUTION

A. Classical Systems

Consider an isolated system of N identical structureless particles. The state of the system at time t can be defined by specifying the phase point $\vec{\Gamma}_t^N = (\vec{p}_t^N, \vec{q}_t^N)$ in the $6N$-dimensional phase space of the system. The state of the system $\vec{\Gamma}_t^N$ changes as the system evolves along a path (trajectory) in phase space in accordance with Hamilton's equations

$$d\vec{q}_j(t)/dt = \partial\mathcal{H}[\,\vec{p}^N(t), \vec{q}^N(t)\,]/\partial\vec{p}_j(t) \tag{198}$$

and

$$d\vec{p}_j(t)/dt = -\partial\mathcal{H}[\,\vec{p}^N(t), \vec{q}^N(t)\,]/\partial\vec{q}_j(t), \tag{199}$$

where $\vec{p}_j(t)$ and $\vec{q}_j(t)$, respectively, are the momentum and coordinate vectors for particle j at time t, and $\mathcal{H}[\,\vec{p}^N(t), \vec{p}^N(t)\,]$ is the classical Hamiltonian for the system. The system is assumed to be conservative. Thus, the Hamiltonian satisfies the condition $\mathcal{H}[\,\vec{p}^N(t), \vec{p}^N(t)\,] = E$, where E is the energy of the system.

In view of the gross nature of macroscopic measurements, we can never specify exactly the state of a real physical system. There will always be some statistical uncertainty in our mesurements. So we adopt a statistical description of classical systems by introducing a time-dependent probability density $\rho(\Gamma^N, t)$, which is defined in such a way that $d^{6N}\Gamma\,\rho(\vec{\Gamma}^N, t)$ represents the probability at time t of finding the system in the phase space-volume element $d^{6N}\Gamma$ in the neighborhood of the phase point $\vec{\Gamma}^N$. Since the phase points of the system must lie in the system phase space, the probability density $\rho(\vec{\Gamma}^N, t)$ is normalized as follows:

$$\int d^{6N}\Gamma\,\rho(\vec{\Gamma}^N, t) = 1. \tag{200}$$

It is convenient to think of the probability density $\rho(\vec{\Gamma}^N, t)$ as describing a fluid occupying the system phase space. Adopting this picture, we have that the $6N$-dimensional vector $\dot{\vec{\Gamma}}^N$ represents the velocity of fluid motion and $\rho(\vec{\Gamma}^N, t)$ represents the fluid density at the point $\vec{\Gamma}^N$. The components of $\vec{\Gamma}^N$ are the collection of components of the three-dimensional vectors $\vec{p}_j$ and $\vec{q}_j$ for each of the particles in the system. Within the context of the fluid picture of $\rho(\vec{\Gamma}^N, t)$, the conservation of probability can be viewed as the conservation of mass. Then $\rho(\vec{\Gamma}^N, t)$ must evolve according to a continuity equation identical in form to the continuity equation for mass density in fluid mechanics. Thus, we write

$$\partial\rho(\vec{\Gamma}^N, t)/\partial t = -\vec{\nabla}_{\Gamma^N} \cdot [\,\dot{\vec{\Gamma}}^N\,\rho(\vec{\Gamma}^N, t)\,], \tag{201}$$

where $\vec{\nabla}_{\Gamma^N}$ is a $6N$-dimensional gradient operator.

The components of the vector $\dot{\vec{\Gamma}}^N$ can be determined by using Hamilton's equations. Computing these vector components in this manner, we obtain $\vec{\nabla}_{\Gamma^N} \cdot \dot{\vec{\Gamma}}^N = 0$. Thus, the continuity equation assumes the form

$$\partial\rho(\vec{\Gamma}^N, t)/\partial t = -\dot{\vec{\Gamma}}^N \cdot \vec{\nabla}_{\Gamma^N}\rho(\vec{\Gamma}^N, t). \tag{202}$$

The partial derivative $\partial\rho(\vec{\Gamma}^N, t)/\partial t$ is the time rate of change of the probability density $\rho(\vec{\Gamma}^N, t)$ at a fixed point in phase space. If we want the time rate of change as seen by an observer moving along a trajectory in phase space, it is necessary to consider the total time derivative

$$d\rho(\vec{\Gamma}^N, t)/dt = \partial\rho(\vec{\Gamma}^N, t)/\partial t + \dot{\vec{\Gamma}}^N \cdot \vec{\nabla}_{\Gamma^N}\rho(\vec{\Gamma}^N, t). \tag{203}$$

It is evident from Eqs. (202) and (203) that the total time derivative $d\rho(\vec{\Gamma}^N, t)/dt$ vanishes. Thus, the probability density $\rho(\vec{\Gamma}^N, t)$ remains constant along a given trajectory in phase space.

Using Hamilton's equations to compute the components of the vector $\dot{\vec{\Gamma}}^N$ in Eq. (202), we obtain the classical Liouville equation

$$\partial\rho(\vec{\Gamma}^N, t)/\partial t = -i\mathcal{L}(\vec{\Gamma}^N)\rho(\vec{\Gamma}^N, t), \tag{204}$$

where $\mathcal{L}(\vec{\Gamma}^N)$ is a differential operator called the Liouville operator. The Liouville operator can be written in the following equivalent forms:

$$\begin{aligned}\mathcal{L}(\vec{\Gamma}^N) &= -i\big[\vec{\nabla}_{p^N}\mathcal{H}(\vec{p}^N, \vec{q}^N) \cdot \vec{\nabla}_{q^N} \\ &\quad - \vec{\nabla}_{q^N}\mathcal{H}(\vec{p}^N, \vec{q}^N) \cdot \vec{\nabla}_{p^N}\big] \qquad (205a) \\ &= -i\sum_{l=1}^{N}\big[\vec{\nabla}_{p_l}\mathcal{H}(\vec{p}^N, \vec{q}^N) \cdot \vec{\nabla}_{q_l} \\ &\quad - \vec{\nabla}_{q_l}\mathcal{H}(\vec{p}^N, \vec{q}^N) \cdot \vec{\nabla}_{p_l}\big] \qquad (205b) \\ &= -i\sum_{l=1}^{N}\left[\frac{\vec{p}_l}{m} \cdot \vec{\nabla}_{q_l} + \vec{F}_l(\vec{q}^N) \cdot \vec{\nabla}_{p_l}\right], \qquad (205c)\end{aligned}$$

where

$$\vec{F}_l(\vec{q}^N) = -\vec{\nabla}_{q_l}U(\vec{q}^N). \tag{206}$$

In the above, $\vec{\nabla}_{p^N}$ and $\vec{\nabla}_{q^N}$, respectively, are $3N$-dimensional momentum and coordinate gradient operators, whereas $\vec{\nabla}_{p_l}$ and $\vec{\nabla}_{q_l}$ are three-dimensional gradient operators associated with particle l. The symbol m in

Eq. (205c) denotes the mass of a given particle. The vector $\vec{F}_l(\vec{q}^N)$, defined by Eq. (206), is the force on particle l arising from the interaction of this particle with the other particles in the system. In writing Eqs. (205c) and (206), we assumed that the classical Hamiltonian $\mathcal{H}(\vec{p}^N, \vec{q}^N)$ can be written in the form

$$\mathcal{H}(\vec{p}^N, \vec{q}^N) = \sum_{l=1}^{N} \frac{\vec{p}_l \cdot \vec{p}_l}{2m} + U(\vec{q}^N), \tag{207}$$

where $U(\vec{q}^N)$ is the interaction between the particles in the system.

The classical Liouville equation is sometimes written

$$\partial\rho(\vec{p}^N, \vec{q}^N; t)/\partial t = -\{\mathcal{H}(\vec{p}^N, \vec{q}^N),\, \rho(\vec{p}^N, \vec{q}^N; t)\}, \tag{208}$$

where the the Poisson bracket $\{\mathcal{H}(\vec{p}^N, \vec{q}^N),\, \rho(\vec{p}^N, \vec{q}^N; t)\}$ is defined by

$$\{\mathcal{H}(\vec{p}^N, \vec{q}^N),\, \rho(\vec{p}^N, \vec{q}^N; t)\} = \mathcal{H}(\vec{p}^N, \vec{q}^N)\Lambda(\vec{p}^N, \vec{q}^N)\rho(\vec{p}^N, \vec{q}^N; t), \tag{209}$$

with $\Lambda(\vec{p}^N, \vec{q}^N)$ denoting the Poisson bracket operator

$$\Lambda(\vec{p}^N, \vec{q}^N) = \vec{\nabla}_{p^N} \cdot \vec{\nabla}_{q^N} - \vec{\nabla}_{q^N} \cdot \vec{\nabla}_{p^N}. \tag{210}$$

The arrows over the gradient operators indicate their direction of operation when the Poisson bracket operator $\Lambda(\vec{p}^N, \vec{q}^N)$ is inserted in Eq. (209).

If the initial probablity density $\rho(\vec{\Gamma}^N, 0)$ is known, we can determine the probability density $\rho(\vec{\Gamma}^N, t)$ at time t by using the formal solution

$$\rho(\vec{\Gamma}^N, t) = \exp[-i\mathcal{L}(\vec{\Gamma}^N)t]\,\rho(\vec{\Gamma}^N, 0) \tag{211}$$

of the classical Liouville equation. The quantity $\exp[-i\mathcal{L}(\vec{\Gamma}^N)t]$ is the propagator for the probability density.

The classical Liouville equation has the following properties: (i) The classical canonical probability density is stationary with respect to the classical Liouville equation. (ii) The classical Liouville operator $\mathcal{L}(\vec{\Gamma}^N)$ is Hermitian. (iii) The classical Liouville equation is invariant under the time-reversal transformation $t \to -t$, $\vec{p}_j \to -\vec{p}_j$, and $\vec{q}_j \to \vec{q}_j$. (iv) The Gibbs entropy $S(t)$ is time independent when $S(t)$ is determined using the formal solution $\rho(\vec{\Gamma}^N, t)$ of the classical Liouville equation.

Properties (ii)–(iv) of the classical Liouville equation are a bit troublesome. The Hermiticity of the classical Liouville operator $\mathcal{L}(\vec{\Gamma}^N)$ implies that its eigenvalues are real. Thus, $\rho(\vec{\Gamma}^N, t)$ must exhibit oscillatory temporal behavior and appears not to decay to a unique stationary state in the limit $t \to \infty$. This raises the question of how do we describe the irreversible decay of a system to a unique equilibrium state. The time-reversal invariance of the classical Liouville equation leads us to the conclusion that this equation describes reversible systems with no privileged direction in time. These problems coupled with the time independence of the Gibbs entropy raise some serious questions about the compatibility of the second law of thermodynamics, the reversibility of the Liouville equation, the use of Gibbs entropy to describe systems out of equilibruim, and the irreversible decay of a system to a unique equilibrium state. This compatibility problem has preoccupied researchers for many years. At this time, there is no satisfactory solution of the compatibility problem.

In principle, the time evolution of the classical dynamical variable $O(\vec{\Gamma}^N, t) = O[\,\vec{p}^N(t), \vec{q}^N(t)\,]$ can be determined by solving Hamilton's equations with the initial conditions $\vec{p}^N(0) = \vec{p}^N$ and $\vec{q}^N(0) = \vec{q}^N$. Alternatively, $O(\vec{\Gamma}^N, t)$ can be determined by using the equation

$$O(\vec{\Gamma}^N, t) = \exp[+i\mathcal{L}(\vec{\Gamma}^N)t]O(\vec{\Gamma}^N), \tag{212}$$

where $\exp[+i\mathcal{L}(\vec{\Gamma}^N)t]$ is the propagator for classical dynamical variables. This equation is the formal solution of the equation of motion

$$dO(\vec{\Gamma}^N, t)/dt = i\mathcal{L}(\vec{\Gamma}^N)O(\vec{\Gamma}^N, t). \tag{213}$$

The average value $\langle O(t)\rangle$ of the classical dynamical variable $O(\vec{\Gamma}^N)$ at time t can be determined by using either of the following relations:

$$\langle O(t)\rangle = \int d^{6N}\Gamma\, \rho(\vec{\Gamma}^N, 0)\, O(\vec{\Gamma}^N, t) \tag{214a}$$

$$= \int d^{6N}\Gamma\, \rho(\vec{\Gamma}^N, t) O(\vec{\Gamma}^N), \tag{214b}$$

where $\rho(\vec{\Gamma}^N, t)$ and $O(\vec{\Gamma}^N, t)$ are given by Eqs. (211) and (212), respectively.

The solution of time evolution problems for classical systems is facilitated by introducing a classical phase space representation that plays a role in the description of classical systems in a manner that is formally analogous to the role played by the coordinate and momentum representations in quantum mechanics. The state vectors $\{|\vec{\Gamma}^N\rangle\}$ of this representation enumerate all of the accessible phase points. The phase function $f(\vec{\Gamma}^N)$ is given by $f(\vec{\Gamma}^N) = \langle\vec{\Gamma}^N | f\rangle$, which can be thought to represent a component of the vector $|f\rangle$ in the classical phase space representation. The application of the classical Liouville operator $\mathcal{L}(\vec{\Gamma}^N)$ to the phase function $f(\vec{\Gamma}^N)$ is defined by $\mathcal{L}(\vec{\Gamma}^N) f(\vec{\Gamma}^N) = \langle\vec{\Gamma}^N|\hat{\mathcal{L}}\,|f\rangle$, where $\hat{\mathcal{L}}$ is an abstract operator that can be associated with the Liouville operator $\mathcal{L}(\vec{\Gamma}^N)$. The inner product $\langle A|B\rangle$ of $\langle A|$ and $|B\rangle$ is defined by $\langle A|B\rangle = \mathrm{Tr}\, A^*B$, where $\mathrm{Tr}\, A^*B$ denotes the classical trace $\mathrm{Tr}\, A^*B = \int d^{6N}\Gamma\, A^*(\vec{\Gamma}^N)B(\vec{\Gamma}^N)$. The closure and orthonormality relations for the classical phase space

representation are given by $\hat{I} = \int d^{6N}\Gamma |\vec{\Gamma}^N\rangle\langle\vec{\Gamma}^N|$ and $\langle\vec{\Gamma}^N|\vec{\Gamma}'^N\rangle = \delta(\vec{\Gamma}^N - \vec{\Gamma}'^N)$, respectively.

With the classical phase space representation at our disposal, we can write the equations of motion for $\rho(\vec{\Gamma}^N, t)$ and $O(\vec{\Gamma}^N, t)$ as

$$\frac{d}{dt}|\rho_t\rangle = -i\hat{\mathcal{L}}|\rho_t\rangle \tag{215}$$

and

$$\frac{d}{dt}|O_t\rangle = i\hat{\mathcal{L}}|O_t\rangle. \tag{216}$$

The formal solutions of these equations can be written

$$|\rho_t\rangle = \exp(-i\hat{\mathcal{L}}t)|\rho_0\rangle \tag{217}$$

and

$$|O_t\rangle = \exp(+i\hat{\mathcal{L}}t)|O\rangle. \tag{218}$$

The results given by Eqs. (217) and (218) enable us to write the average value $\langle O(t)\rangle$ of the classical dynamical variable $O(\vec{\Gamma}^N)$ as

$$\langle O(t)\rangle = \left\langle O_t^* \middle| \rho_0 \right\rangle \tag{219a}$$

$$= \left\langle O^* \middle| \rho_t \right\rangle \tag{219b}$$

$$= \langle O^*| \exp(-i\hat{\mathcal{L}}t)|\rho_0\rangle, \tag{219c}$$

where the asterisk indicates complex conjugation.

B. Quantum Systems

In quantum mechanics, the state of a system at time t is defined by specifying a state vector $|\psi_t\rangle$. The time evolution of $|\psi_t\rangle$ is governed by the Schrödinger equation

$$\frac{d}{dt}|\psi_t\rangle = -\left(\frac{i}{\hbar}\right)\hat{\mathcal{H}}|\psi_t\rangle, \tag{220}$$

where $\hat{\mathcal{H}}$ is the quantum Hamiltonian for the system.

In principle, we can determine the state vector $|\psi_t\rangle$ from the initial state vector $|\psi_0\rangle$ by using the formal solution

$$|\psi_t\rangle = \exp\left[-\left(\frac{i}{\hbar}\right)\hat{\mathcal{H}}t\right]|\psi_0\rangle \tag{221}$$

of the Schrödinger equation, where $\exp[-(i/\hbar)\hat{\mathcal{H}}t]$ is the propagator for state vectors.

The application of Eq. (221) to the description of the time evolution of a quantum system requires us to specify the initial state of the system. Generally, there is some statistical indeterminancy in its initial preparation. Thus, we adopt a statistical description that employs the density operator $\hat{\rho}(0)$ to specify the initial state. As with equilibrium density operators, the density operator $\hat{\rho}(0)$ is assumed to have the following properties: (i) $\mathrm{Tr}\,\hat{\rho}(0) = 1$, (ii) $\hat{\rho}(0)$ is Hermitian, and (iii) the diagonal matrix elements of $\hat{\rho}(0)$ in any representation are nonnegative and represent the probabilities of finding the system in the various states of that representation.

Given the eigenvectors $\{|\psi\rangle\}$ and eigenvalues $\{P(\psi)\}$ of the density operator $\hat{\rho}(0)$, we can determine the probability $P(\alpha; t)$ of finding the system at time t in the eigenstate $|\alpha\rangle$ of the Hermitian operator $\hat{A}$ by using the relation

$$P(\alpha;\, t) = \sum_{\psi} P(\alpha|\psi; t)P(\psi). \tag{222}$$

$P(\psi)$ represents the probability of finding the system at time $t = 0$ in the eigenstate $|\psi\rangle$ of $\hat{\rho}(0)$. The quantity $P(\alpha \mid \psi; t) = |\langle\alpha|\psi_t\rangle|^2$ is a conditional probability. It represents the probability of the system making a transition from the state $|\psi\rangle$ to the state $|\alpha\rangle$ during the time interval t when the system has been prepared in the state $|\psi\rangle$ at time $t = 0$. $P(\psi)$ is due to the lack of initial information, whereas $P(\alpha \mid \psi; t)$ is due to the statistical nature of quantum mechanics.

Making use of Eq. (222), we find that the probability $P(\alpha; t)$ can be written as the diagonal matrix element $\langle\alpha|\hat{\rho}(t)|\alpha\rangle$ of the time-dependent density operator

$$\hat{\rho}(t) = \exp\left[-\left(\frac{i}{\hbar}\right)\hat{\mathcal{H}}t\right]\hat{\rho}(0)\exp\left[+\left(\frac{i}{\hbar}\right)\hat{\mathcal{H}}t\right], \tag{223}$$

where

$$\hat{\rho}(0) = \sum_{\psi} |\psi\rangle P(\psi)\langle\psi|. \tag{224}$$

Although the ψ-representation was used to arrive at Eq. (223), the initial density operator $\hat{\rho}(0)$ need not be expressed in the ψ-representation. In actual practice, one usually chooses a representation on the basis of mathematical convenience for the problem at hand.

As indicated above, the diagonal matrix elements $\rho(\alpha, \alpha; t)$ of the density operator $\hat{\rho}(t)$ in the α-representation give us the probabilities of finding the system at time t in the various states of that representation. The off-diagonal matrix elements $\rho(\alpha, \alpha'; t)$ of $\hat{\rho}(t)$ provide us with information about the phase coherence between these states at time t. This interpretation of the matrix elements of the density operator apples to any representation for which the matrix elements of $\hat{\rho}(t)$ are defined.

The time derivative of Eq. (223) gives us the following equation of motion for the density operator $\hat{\rho}(t)$:

$$\frac{d}{dt}\hat{\rho}(t) = -\left(\frac{i}{\hbar}\right)[\hat{\mathcal{H}}, \hat{\rho}(t)]_-. \tag{225}$$

The subscript minus sign in $[\hat{\mathcal{H}}, \hat{\rho}(t)]_-$ indicates that this quantity is a commutator. Equation (225) is the quantum analogue of the classical Liouville equation. Thus, we call it the quantum Liouville equation.

The quantum Liouville equation can be brought into a form that more closely resembles the classical Liouville equation by introducing the quantum Liouville operator

$$\hat{\hat{\mathcal{L}}} = \left(\frac{i}{\hbar}\right)\hat{\mathcal{H}}^{-}, \tag{226}$$

where $\hat{\hat{\mathcal{H}}}^{-}$ is defined in such a way that

$$\hat{\hat{\mathcal{H}}}^{-}\hat{A} = [\hat{\mathcal{H}}, \hat{A}]_{-} \tag{227}$$

for any operator $\hat{A}$. Making use of Eqs. (226) and (227), we rewrite Eq. (225) as

$$\frac{d}{dt}\hat{\rho}(t) = -i\hat{\hat{\mathcal{L}}}\hat{\rho}(t). \tag{228}$$

The double caret in the quantities $\hat{\hat{\mathcal{L}}}$ and $\hat{\hat{\mathcal{H}}}^{-}$ indicates that they are tetradic operators, i.e., operators that require four indices in their matrix representation. Such operators are sometimes called superoperators.

If the initial density operator $\hat{\rho}(0)$ is known, the density operator $\hat{\rho}(t)$ at time t can be determined by using Eq. (223). This equation is the formal solution of the version of the quantum Liouville equation given by Eq. (225). In view of the equivalence of Eqs. (225) and (228), we can also write

$$\hat{\rho}(t) = \exp(-i\hat{\hat{\mathcal{L}}}t)\hat{\rho}(0). \tag{229}$$

This is the formal solution of Eq. (228). Clearly, we must have

$$\exp(-i\hat{\hat{\mathcal{L}}}t)\hat{\rho}(0) = \exp\left[-\left(\frac{i}{\hbar}\right)\hat{\mathcal{H}}t\right]\hat{\rho}(0)\exp\left[+\left(\frac{i}{\hbar}\right)\hat{\mathcal{H}}t\right]. \tag{230}$$

The basic properties of the classical Liouville equation and the troublesome questions they raise are shared by the quantum Liouville equation. For the quantum case, we summarize these properties as follows: (i) The canonical density operator is stationary with respect to the quantum Liouville equation. (ii) The quantum Liouville operator $\hat{\hat{\mathcal{L}}}$ is Hermitian. (iii) The quantum Liouville equation is time-reversal invariant. (iv) The Gibbs entropy $S(t)$ is time independent when $S(t)$ is determined using the formal solution $\hat{\rho}(t)$ of the quantum Liouville equation.

Given the density operator $\hat{\rho}(t)$, we can determine the average value $\langle\hat{O}(t)\rangle$ of the quantum dynamical variable $\hat{O}$ at time t by using the relation

$$\langle\hat{O}(t)\rangle = \mathrm{Tr}\,\hat{\rho}(t)\hat{O}. \tag{231}$$

This corresponds to the Schrödinger picture of quantum mechanics.

Alternatively, we can determine $\langle\hat{O}(t)\rangle$ by working in the Heisenberg picture. For this picture,

$$\langle\hat{O}(t)\rangle = \mathrm{Tr}\,\hat{\rho}(0)\hat{O}(t), \tag{232}$$

where

$$\hat{O}(t) = \exp\left[+\left(\frac{i}{\hbar}\right)\hat{\mathcal{H}}t\right]\hat{O}\exp\left[-\left(\frac{i}{\hbar}\right)\hat{\mathcal{H}}t\right]. \tag{233}$$

If the system is initially prepared in some state $|\psi_0\rangle$, the initial density operator $\hat{\rho}(0)$ is given by $\hat{\rho}(0) = |\psi_0\rangle\langle\psi_0|$. For this case, Eqs. (231) and (232) reduce to the forms usually employed in quantum mechanics, i.e.,

$$\langle\hat{O}(t)\rangle = \langle\psi_t|\hat{O}|\psi_t\rangle \tag{234}$$

for the Schrödinger picture and

$$\langle\hat{O}(t)\rangle = \langle\psi_0|\hat{O}(t)|\psi_0\rangle \tag{235}$$

for the Heisenberg picture.

The time evolution of the quantum dynamical variable $\hat{O}(t)$ can be generated by using Eq. (233). This result is the formal solution of Heisenberg's equation of motion

$$\frac{d}{dt}\hat{O}(t) = \left(\frac{i}{\hbar}\right)[\hat{\mathcal{H}}, \hat{O}(t)]_{-}. \tag{236}$$

Introducing the formal definition of the quantum Liouville operator $\hat{\hat{\mathcal{L}}}$, we can cast Eq. (236) in a form that resembles the equation of motion for classical dynamical variables. More specifically, we can write

$$\frac{d}{dt}\hat{O}(t) = i\hat{\hat{\mathcal{L}}}\hat{O}(t). \tag{237}$$

In view of the formal equivalence of Eqs. (236) and (237), $\hat{O}(t)$ is also given by

$$\hat{O}(t) = \exp(+i\hat{\hat{\mathcal{L}}}t)\,\hat{O}. \tag{238}$$

This result is the formal solution of Eq. (237).

Matrix representations of the quantum Liouville equation and Heisenberg's equation of motion can be obtained by sandwiching both sides of Eqs. (225) and (236) or Eqs. (228) and (237) between the vectors $\langle\phi_j|$ and $|\phi_k\rangle$, where $|\phi_j\rangle$ and $|\phi_k\rangle$ are members of the orthonormal basis $\{|\phi_l\rangle\}$. This procedure yields

$$\frac{\partial}{\partial t}\rho(j,k;t) = -i\sum_{l,m}\mathcal{L}(jk,lm)\rho(l,m;t) \tag{239}$$

and

$$\frac{d}{dt}O(j,k;t) = i\sum_{l,m}\mathcal{L}(jk,lm)O(l,m;t), \tag{240}$$

where

$$\rho(j,k;t) = \langle\phi_j|\hat{\rho}(t)|\phi_k\rangle, \qquad O(j,k;t) = \langle\phi_j|\hat{O}(t)|\phi_k\rangle, \tag{241}$$

and

$$\mathcal{L}(jk,lm) = \left(\frac{i}{\hbar}\right)[\mathcal{H}(j,l)\delta_{m,k} - \mathcal{H}(m,k)\delta_{j,l}], \tag{242}$$

with

$$\mathcal{H}(j,l) = \langle\phi_j|\hat{\mathcal{H}}|\phi_l\rangle. \tag{243}$$

Although the above matrix representations of the quantum Liouville equation and Heisenberg's equation of motion are formally correct, the solution of time evolution problems for quantum systems can be more readily accomplished by working in a representation called the superstate representation. The basis vectors of this representation are the superstates $\{|N_{jk}\rangle\}$. These states are associated with the operators $\{\hat{N}_{jk} = |\phi_j\rangle\langle\phi_k|\}$ formed from the basis vectors $\{|\phi_j\rangle\}$ used in the formulation of Eqs. (239) and (240). The matrix element $A(j,k) = \langle\phi_j|\hat{A}|\phi_k\rangle$ of the operator $\hat{A}$ is given by $A(j,k) = \langle N_{jk}|A\rangle$, which can be thought to represent a component of the vector $|A\rangle$ in the superstate representation. We define the inner product $\langle A|B\rangle$ of $\langle A|$ and $|B\rangle$ by $\langle A|B\rangle = \mathrm{Tr}\,\hat{A}^\dagger\hat{B}$. The matrix elements $\mathcal{L}(jk,lm)$ of the quantum Liouville operator $\hat{\hat{\mathcal{L}}}$ are given by $\mathcal{L}(jk,lm) = \langle N_{jk}|\hat{\mathcal{L}}|N_{lm}\rangle$, where $\hat{\mathcal{L}}$ is an abstract operator that can be associated with the Liouville operator $\hat{\hat{\mathcal{L}}}$. The closure and orthonormality relations for the superstate representation are given by $\hat{I} = \sum_{j,k}|N_{jk}\rangle\langle N_{jk}|$ and $\langle N_{jk}|N_{lm}\rangle = \delta_{j,l}\delta_{k,m}$, respectively.

With the superstate representation at our disposal, we can rewrite the quantum Liouville equation and Heisenberg's equation of motion as vector equations of motion that are identical in form to the vector equations of motion given by Eqs. (215) and (216) for classical systems. The only difference between the quantum and classical vector equations of motion is the manner in which the matrix elements of $\hat{\mathcal{L}}$ and the components of $|O_t\rangle$ and $|\rho_t\rangle$ are determined. Nonetheless, the expressions for the average of a quantum dynamical variable differ from the corresponding expressions for classical systems. [See Eqs. (219a)–(219c).] For the quantum case, we have

$$\langle\hat{O}(t)\rangle = \langle O^\dagger|\rho_t\rangle \tag{244a}$$

$$= \langle O_t^\dagger|\rho_0\rangle \tag{244b}$$

$$= \langle O^\dagger|\exp(-\hat{L}t)|\rho_0\rangle, \tag{244c}$$

where the dagger indicates the Hermitian conjugate.

V. PHENOMENOLOGICAL DESCRIPTIONS OF NONEQUILIBRIUM SYSTEMS

A. Phenomenological Equations of Motion

The quantum and classical Liouville equations are rarely used in the actual characterization of experimental data concerning the spectral and temporal properties of real physical systems. Instead, investigators usually adopt a contracted description of the physical system under consideration. Such a description entails the use of an equation of motion for a density operator or probability density characterizing only the relevant part of the physical system. Much of the underlying dynamics is buried in parameters intended to describe damping arising from the interaction between the relevant and irrelevant parts of the system. Equations of motion endowed with these features are often referred to as phenomenological equations.

The most commonly used phenomenological equations are linear equations of motion that can be cast in a vector form that is identical to Eq. (215) with the operator $\hat{\mathcal{L}}$ playing the role of an effective Liouville operator. Unlike the Liouville equations discussed earlier, these linear phenomenological equations possess broken time-reversal symmetry. Moreover, the effective Liouville operator in such equations is non-Hermitian. In view of this property, a system can display damped oscillatory behavior, resulting in the decay of an initially prepared nonequilibrium state. Often this decay is such that the system tends to evolve toward a stationary state (usually thermal equilibrium) as $t \to \infty$.

Linear equations of motion possessing the above-described properties include a host of equations commonly referred to as master equations, Fokker–Planck equations, and stochastic Liouville equations. Many of the equations belonging to these classes of dynamical models were originally developed on the basis of intuitive arguments about the nature of a physical system. For the most part, phenomenological equations have been quite successful in the codification of large amounts of experimental data. This success has generated much theoretical work concerned with the construction of phenomenological equations from basic principles, thus affording us with formal expressions that relate phenomenological parameters to microscopic interactions. In Section 6, we discuss some of the approaches that have been employed to obtain such expressions.

Apart from constructing phenomenological equations and formal expressions for phenomenological parameters, researchers have focused much attention on the construction of new phenomenological equations, studying the range of validity of known phenomenological equations and generalizing such equations to include

non-Markovian retardation and nonlinearities. In addition, considerable effort has been made to develop powerful techniques that enable us to compute physically relevant quantities without having to obtain solutions to either global or contracted equations of motion.

In the treatment of the dynamics of N identical classical particles executing motion in a spatially homogeneous environment with a fixed temperature T, investigators often adopt the N-particle classical Fokker–Planck equation

$$\partial\rho(\vec{\Gamma}^N,t)/\partial t = \left[-i\mathcal{L}_S(\vec{\Gamma}^N) + \mathbf{L}_{FP}(\vec{\Gamma}^N)\right]\rho(\vec{\Gamma}^N,t), \tag{245}$$

where

$$\mathcal{L}_S(\vec{\Gamma}^N) = -i\sum_{j=1}^{N}\left[\frac{\vec{p}_j}{m}\cdot\vec{\nabla}_{q_j} + \vec{\bar{F}}_j(\vec{q}^N)\cdot\vec{\nabla}_{p_j}\right] \tag{246}$$

and

$$\mathbf{L}_{FP}(\vec{\Gamma}^N) = \sum_{j,k=1}^{N}\vec{\nabla}_{p_j}\cdot\vec{\vec{\xi}}_{j,k}\cdot[\vec{p}_k + mk_BT\vec{\nabla}_{p_k}]. \tag{247}$$

In the above, $\vec{\bar{F}}_j(\vec{q}^N) = -\vec{\nabla}_{q_j}\bar{U}(\vec{q}^N)$ denotes the mean force experienced by particle j due to the mean potential $\bar{U}(\vec{q}^N)$ obtained by averaging the total interaction potential for the particles, including the particles in the environment, over the equilibrium probability density of the environment. The quantity $\vec{\vec{\xi}}_{j,k}$ is called the friction tensor. The streaming operator $\mathcal{L}_S(\vec{\Gamma}^N)$ describes the reversible motion of N particles in an environment at thermal equilibrium. Damping of the reversible motion is brought about by the Fokker–Planck operator $\mathbf{L}_{FP}(\vec{\Gamma}^N)$, which leads to the damping of the momentum degrees of freedom. This damping is communicated to the spatial degrees of freedom through the coupling terms $(\vec{p}_l/m)\cdot\vec{\nabla}_{q_l}$ and $\vec{\bar{F}}_l(\vec{q}^N)\cdot\vec{\nabla}_{p_l}$ in $\mathcal{L}_S(\vec{\Gamma}^N)$. In essence, the momentum relaxation drives the relaxation of the spatial degrees of freedom. The solution of the Fokker–Planck equation assumes the form of the classical equilibrium canonical probability density as $t\to\infty$.

On time scales that are long compared with the time scale for momentum relaxation, it is thought that the momentum degrees of freedom are essentially in thermal equilibrium, while the spatial degrees of freedom are significantly out of equilibrium. For such cases, the Fokker–Planck equation is assumed to reduce to the Smoluchowski equation

$$\frac{\partial\rho(\vec{q}^N,t)}{\partial t} = \sum_{j,k=1}^{N}\vec{\nabla}_{q_j}\cdot\vec{\vec{D}}_{jk}\cdot\left[\vec{\nabla}_{q_k} - \frac{1}{k_BT}\vec{\bar{F}}_k(\vec{q}^N)\right]\rho(\vec{q}^N,t), \tag{248}$$

where $\rho(\vec{q}^N,t)$ is a spatial probability density and $\vec{\vec{D}}_{jk}(\vec{q}^N)$ is called the diffusion tensor. In the infinite-time limit $t\to\infty$, the solution of the Smoluchowski equation assumes the form of the classical equilibrium canonical spatial probability density $\exp[\Omega_S, -(1/k_BT)\bar{U}(\vec{q}^N)]$. The Smoluchowski and Fokker–Planck equations have been used to describe a diversity of phenomena, including coagulation, dynamics of colloidal systems, electrolytic processes, chemical reactions, ion transport in biological systems, diffusion of particles on surfaces, and sedimentation.

In treating the excited-state dynamics of a collection of atoms/molecules interacting with an environment held at fixed temperature T, investigators often adopt a generalized master equation of the form

$$\partial\rho(j,k;t)/\partial t = -\sum_{l,m}[i\mathcal{L}_S(jk,lm) + R(jk,lm)]\rho(l,m;t), \tag{249}$$

where

$$\mathcal{L}_S(jk,lm) = \left(\frac{i}{\hbar}\right)[\bar{\mathcal{H}}(j,l)\delta_{m,k} - \bar{\mathcal{H}}(m,k)\delta_{j,l}]. \tag{250}$$

In the above, the matrix elements $\bar{\mathcal{H}}(j,l) = \langle\phi_j|\hat{\bar{\mathcal{H}}}|\phi_l\rangle$ are those of a mean Hamiltonian $\hat{\bar{\mathcal{H}}}$ obtained by averaging the Hamiltonian for the entire system, including the environment, over the equilibrium density operator for the environment. The matrix elements $\mathcal{L}_S(jk,lm)$ of the tetradic operator $\hat{\hat{\mathcal{L}}}_S$ describe the reversible motion of the collection of atoms/molecules in an environment at thermal equlibrium. Damping of the reversible motion is brought about by processes described by the matrix elements $R(jk,lm)$ of the relaxation tetradic $\hat{\hat{R}}$. These matrix elements are usually defined in such a way that the collection of atoms/molecules achieves thermal equilibrium as $t\to\infty$.

It is generally believed that the relaxation of the phase coherence (dephasing) occurs on a shorter time scale than the time scale for the decay of excited state populations. For times much longer than the time scale for the loss of phase coherence, the generalized master equation is assumed to take the form of the kinetic equation

$$\partial\rho(j,j;t)/\partial t = -\rho(j,j;t)W(j\to) + \sum_{k\neq j}\rho(k,k;t)W(k\to j), \tag{251}$$

where $W(k\to j)$ is a rate constant for the transition from the state $|\phi_k\rangle$ to the state $|\phi_j\rangle$, and $W(j\to) = \sum_{k\neq j}W(j\to k)$ is the total rate constant for transitions out of the state $|\phi_j\rangle$. The rate constants are usually required to satisfy the principle of detailed balance, i.e., $\rho_{eq}(j,j)W(j\to k) = \rho_{eq}(k,k)W(k\to j)$, in order for the

excited-state populations $\{\rho(j,\, j;t)\}$ to decay to the equilibrium values $\{\rho_{\rm eq}(j,\, j)\}$ as $t \to \infty$.

In some dynamical models, the temporal evolutions of the diagonal and off-diagonal matrix elements of the density operator are decoupled. For such models, the temporal evolution of the diagonal matrix elements is generally assumed to be given by Eq. (251) for all times.

B. Classical Brownian Motion Theory

Around the turn of the century researchers in the area of Brownian motion theory were preoccupied with the irregular motion exhibited by colloidal-sized particles immersed in a fluid. Since then the mathematical apparatus of Brownian theory has crept into a number of disciplines and has been used to treat a range of problems involving systems from the size of atoms to systems of stellar dimensions. For the sake of clarity, we shall not consider such a range of problems, confining our attention to the more traditional problem of describing a collection of N identical classical particles executing motion in a thermal environment.

In classical Brownian motion theory, one usually assumes that the time evolution of the probability density $\rho(\vec{\Gamma}^N, t)$ is governed by the Chapman–Kolmogorov equation

$$\rho(\vec{\Gamma}^N, t+\Delta t) = \int d^{6N}\Gamma' P(\vec{\Gamma}^N \mid \vec{\Gamma}'^N; \Delta t)\rho(\vec{\Gamma}'^N, t), \tag{252}$$

where $P(\vec{\Gamma}^N \mid \vec{\Gamma}'^N; \Delta t)$ is the probability of the N-particle system making a transition from the phase point $\vec{\Gamma}'^N$ to the phase point $\vec{\Gamma}^N$ during the time interval Δt. Self-consistency requires the conditional transition probabilities $P(\vec{\Gamma}^N \mid \vec{\Gamma}'^N; \Delta t)$ to be normalized to unity, i.e., $\int d^{6N}\Gamma\, P(\vec{\Gamma}^N \mid \vec{\Gamma}'^N; \Delta t) = 1$, and to satisfy the boundary condition $P(\vec{\Gamma}^N \mid \vec{\Gamma}'^N; \Delta t = 0) = \delta(\vec{\Gamma}^N - \vec{\Gamma}'^N)$.

The time interval Δt is assumed to represent the time scale of macroscopic measurements, i.e., the time resolution of the observer. It is also assumed that Δt is short on the time scale t_S characterizing the motion of the N particles (the system) and long on the time scale t_B characterizing the response of the environment (the bath) to the motion of the system and its re-equilibration. This assumption requires the existence of a separation in the time scales t_S and t_B for the evolution of the system and bath, the latter being faster ($t_B \ll t_S$), so that $t_B \ll \Delta t \ll t_S$.

It is convenient to introduce a rate constant $W(\vec{\Gamma}'^N \to \vec{\Gamma}^N; \Delta t)$ characterizing the average rate at which the N-particle system passes from the phase point $\vec{\Gamma}'^N$ to the phase point $\vec{\Gamma}^N$ during the time interval Δt. We define $W(\vec{\Gamma}'^N \to \vec{\Gamma}^N; \Delta t)$ by

$$P(\vec{\Gamma}^N \mid \vec{\Gamma}'^N; \Delta t) = \delta(\vec{\Gamma}^N - \vec{\Gamma}'^N) + \Delta t\, W(\vec{\Gamma}'^N \to \vec{\Gamma}^N; \Delta t). \tag{253}$$

With this definition, we can readily convert the Chapman–Kolmogorov equation to a phase space master equation

$$\partial\rho(\vec{\Gamma}^N, t; \Delta t)/\partial t = \int d^{6N}\Gamma' W(\vec{\Gamma}'^N \to \vec{\Gamma}; \Delta t)\rho(\vec{\Gamma}'^N, t), \tag{254}$$

where $\partial\rho(\vec{\Gamma}^N, t; \Delta t)/\partial t$ is a phenomenological time derivative defined by

$$\partial\rho(\vec{\Gamma}^N, t; \Delta t)/\partial t = [\rho(\vec{\Gamma}^N, t+\Delta t) - \rho(\vec{\Gamma}^N, t)]/\Delta t \tag{255a}$$

$$= \Delta t^{-1}\int_t^{t+\Delta t} dt'\partial\rho(\vec{\Gamma}^N, t')/\partial t'. \tag{255b}$$

The phenomenological time derivative is a coarse-grained time derivative obtained by time averaging the instantaneous time derivative $\partial\rho(\vec{\Gamma}^N, t')/\partial t'$ over the time scale Δt of macroscopic measurements. From a macroscopic point of view, the phenomenological time derivative can be treated as an instantaneous time derivative. In adopting this attitude, one is mimicking the kind of time smoothing actually done in the analysis of experiments, which are always coarse-grained in time.

By performing a derivate moment expansion of the rate constants appearing in the phase space master equation, one can convert this integral equation to an equivalent differential equation called the generalized Fokker–Planck equation:

$$\frac{\partial\rho(\vec{\Gamma}^N, t; \Delta t)}{\partial t} = \sum_{s_1=0}^{\infty}\cdots\sum_{s_{6N}=0}^{\infty}\left[\frac{(-1)^{(s_1+\cdots+s_{6N})}}{s_1!\cdots s_{6N}!}\right]\nabla_{\Gamma_1}^{s_1}\cdots \times \nabla_{\Gamma_{6N}}^{s_{6N}}\left[\mathbb{K}_{\Gamma_1,\ldots,\Gamma_{6N}}^{(s_1+\cdots+s_{6N})}(\vec{\Gamma}^N)\rho(\vec{\Gamma}^N, t)\right], \tag{256}$$

where the derivate moments $\mathbb{K}_{\Gamma_1,\ldots,\Gamma_{6N}}^{(s_1+\cdots+s_{6N})}(\vec{\Gamma}^N)$ are defined by

$$\mathbb{K}_{\Gamma_1,\ldots,\Gamma_{6N}}^{(s_1+\cdots+s_{6N})}(\vec{\Gamma}^N) = \int d^{6N}\Gamma'(\Gamma_1' - \Gamma_1)^{s_1}\cdots(\Gamma_{6N}' - \Gamma_{6N})^{s_{6N}} \times W(\vec{\Gamma}^N \to \vec{\Gamma}'^N; \Delta t), \tag{257}$$

with Γ_l denoting a component of the $6N$-dimensional vector $\vec{\Gamma}^N$. The derivate moments characterize the distribution of spatial and momentum transitions in phase space.

If the derivate moments higher than second order are neglected, Eq. (256) assumes the form of a so-called linear Fokker–Planck equation:

$$\partial\rho(\vec{\Gamma}^N, t; \Delta t)/\partial t = -\vec{\nabla}_{\Gamma^N} \cdot \left[\vec{\mathbb{K}}^{(1)}_{\Gamma^N}(\vec{\Gamma}^N)\rho(\vec{\Gamma}^N, t)\right] + \frac{1}{2}\vec{\vec{\nabla}}_{\Gamma^N} : \left[\vec{\vec{\mathbb{K}}}^{(2)}_{\Gamma^N,\Gamma^N}(\vec{\Gamma}^N)\rho(\vec{\Gamma}^N, t)\right], \tag{258}$$

where $\vec{\nabla}_{\Gamma^N}$ is a $6N$-dimensional gradient operator. The components

$$\mathbb{K}^{(1)}_{\Gamma_i}(\vec{\Gamma}^N) = \int d^{6N}\Gamma'(\Gamma'_i - \Gamma_i)\left[\frac{P(\vec{\Gamma}'^N \mid \vec{\Gamma}^N; \Delta t)}{\Delta t}\right] \tag{259}$$

of $\vec{\mathbb{K}}^{(1)}_{\Gamma^N}(\vec{\Gamma}^N)$ are called "drift" coefficients and the components

$$\mathbb{K}^{(2)}_{\Gamma_i,\Gamma_j}(\vec{\Gamma}^N) = \int d^{6N}\Gamma'(\Gamma'_i - \Gamma_i)(\Gamma'_j - \Gamma_j) \times \left[\frac{P(\vec{\Gamma}'^N \mid \vec{\Gamma}^N; \Delta t)}{\Delta t}\right] \tag{260}$$

of $\vec{\vec{\mathbb{K}}}^{(2)}_{\Gamma^N,\Gamma^N}(\vec{\Gamma}^N)$ are called "diffusion" coefficients. The neglect of derivate moments higher than second order is equivalent to assuming that the spatial and momentum transitions are short ranged.

At this point, it is usually assumed that the average of the classical variables over the conditional transition probabilities, so-called stochastic averaging, is equivalent to averaging the classical equations of motion over the equilibrium distribution of the bath. More specifically,

$$\mathbb{K}^{(1)}_{\Gamma_i}(\vec{\Gamma}^N) = \langle[\Gamma_i(\Delta t) - \Gamma_i(0)]\rangle_B/\Delta t \tag{261}$$

and

$$\mathbb{K}^{(2)}_{\Gamma_i,\Gamma_j}(\vec{\Gamma}^N) = \langle[\Gamma_i(\Delta t) - \Gamma_i(0)][\Gamma_j(\Delta t) - \Gamma_j(0)]\rangle_B/\Delta t, \tag{262}$$

where $\langle\,\rangle_B$ denotes averaging over the equilibrium distribution of the bath.

The actual evaluation of Eqs. (261) and (262) requires us to solve the classical equations of motion for the N particles plus bath. This many-body problem is circumvented by introducing a stochastic description of the N-particle motion with the bath treated as a source of thermal noise. This is accomplished by introducing Langevin equations, which correspond to a mean field version of Hamilton's equations for the N particles augmented with additional stochastic terms intended to represent the influence of the bath on the N-particle motion.

For the case of a single particle of mass m in a spatially homogeneous bath, we write the Langevin equations as

$$d\vec{p}(t)/dt = -\xi\,\vec{p}(t) + \vec{f}(t) \tag{263}$$

and

$$d\vec{q}(t)/dt = \vec{p}(t)/m. \tag{264}$$

The first term on the right side of Eq. (263) is a frictional force due to the viscous drag exerted on the particle by the bath. The quantity $\vec{f}(t)$ appearing in the second term is a fluctuating force intended to represent the influence of the collisions between the particle of interest and the particles of the bath.

The assumption that the frictional force is proportional to the particle's momentum $\vec{p}$ is motivated by Stoke's law. According to Stoke's law, the frictional force on a spherical particle of mass m and radius r in a medium of viscosity η is $-\xi\,\vec{p}$, where $\xi = 6\pi r\eta/m$ is called the friction coefficient.

The force $\vec{f}(t)$ is assumed to be a random force that fluctuates on time scales much smaller than the time scale characterizing the motion of the momentum $\vec{p}(t)$. The random nature of $\vec{f}(t)$ is expressed by

$$\langle f_\alpha(t)\rangle_B = 0, \tag{265}$$

where the index α indicates a Cartesian component of $\vec{f}(t)$. Since the force $\vec{f}(t)$ is assumed to vary on a much shorter time scale than the momentum $\vec{p}(t)$, the time evolution of $\vec{p}(t)$ and $\vec{f}(t)$ are expected to be uncorrelated, i.e., $p_{\alpha_1}(t_1)$ and $f_{\alpha_2}(t_2)$ are independent of each other. Thus, we write

$$\left\langle p_{\alpha_1}(t_1)\, f_{\alpha_2}(t_2)\right\rangle_B = 0. \tag{266}$$

In addition to the above assumptions, the force $\vec{f}(t)$ is assumed to be δ-correlated in time and Gaussian, i.e.,

$$\left\langle f_{\alpha_1}(t_1)\, f_{\alpha_2}(t_2)\right\rangle_B = 2mk_BT\xi\,\delta_{\alpha_1,\alpha_2}\delta(t_1 - t_2), \tag{267}$$

$$\left\langle f_{\alpha_1}(t_1)\, f_{\alpha_2}(t_2)\cdots f_{\alpha_{2n+1}}(t_{2n+1})\right\rangle_B = 0, \tag{268}$$

and

$$\left\langle f_{\alpha_1}(t_1)\, f_{\alpha_2}(t_2)\cdots f_{\alpha_{2n}}(t_{2n})\right\rangle_B = \sum_{\text{all pairs}} \left\langle f_{\alpha_p}(t_p)\, f_{\alpha_q}(t_q)\right\rangle_B \left\langle f_{\alpha_r}(t_r)\, f_{\alpha_s}(t_s)\right\rangle_B \cdots, \tag{269}$$

where the sum has to be taken over all of the different ways that one can divide the $2n$ time points $t_1, \ldots, t_{2n}$ into n pairs.

Note that the right side of Eq. (267) scales as the temperature T of the bath. The motivation for this scaling and endowing the force $\vec{f}(t)$ with the above-described properties is to bias the dynamics in such a way that the average kinetic energy of the particle $[1/(2m)]\langle\vec{p}(t)\cdot\vec{p}(t)\rangle$ decays to

the equipartition value $3k_BT/2$ as $t \to \infty$, in other words, so that the particle of interest will come to thermal equilibrium with the bath.

The formal solution of the Langevin equations yields

$$\vec{p}(t) = \psi(\xi;t)\vec{p}(0) + \int_0^t d\tau \psi(\xi;t-\tau)\vec{f}(\tau) \tag{270}$$

and

$$\vec{q}(t) = \vec{q}(0) + \left(\frac{1}{m\xi}\right)\vec{p}(0)[1-\psi(\xi;t)] + \left(\frac{1}{m\xi}\right)\int_0^t d\tau[1-\psi(\xi;t-\tau)]\vec{f}(\tau), \tag{271}$$

where

$$\psi(\xi;t) = \exp(-\xi t). \tag{272}$$

With these results and the above assumptions about the nature of $\vec{f}(t)$, we can determine the "drift" and "diffusion" coefficients given by Eqs. (261) and (262). Before doing this, it is instructive to make use of the above results to explore some additional features of Langevin dynamics.

Averaging the formal solution for $\vec{p}(t)$ given by Eq. (270) over the initial probability density for the particle plus bath, assuming the bath is at thermal equilibrium, we find that the average values of the Cartesian components of the initial average momentum $\langle\vec{p}(0)\rangle$ decay exponentially to zero with the same relaxation time $\tau = \xi^{-1}$. Thus, we think of the inverse of the friction coefficient ξ as providing a measure of the time scale for momentum relaxation.

After the particle's initial momentum $\langle\vec{p}(0)\rangle$ has come to thermal equilibrium, there will be spontaneous equilibrium momentum fluctuations. Such fluctuations are described by the equilibrium time correlation function

$$C_{pp}(t) = \langle\vec{p}(t)\cdot\vec{p}(0)\rangle_{\text{eq}}, \tag{273}$$

where $\langle\,\rangle_{\text{eq}}$ denotes averaging over the equilibrium distribution for the particle plus bath. The time evolution of the time correlation function can be determined by using Eq. (270). The result is

$$C_{pp}(t) = \exp(-\xi t)C_{pp}(0). \tag{274}$$

The relaxation time $\tau_C = \xi^{-1}$ provides a measure of the time scale for which the momentum fluctuations are correlated. So we call τ_C the correlation time.

We can also examine spontaneous equilibrium fluctuations by working in the frequency domain and considering the spectral density (Fourier–Laplace transform) of equilibrium time correlation functions. For the case of momentum fluctuations, we write

$$\mathcal{C}_{pp}(i\omega+\epsilon) = \int_0^\infty dt\,\exp[-(i\omega+\epsilon)t]\,C_{pp}(t) \tag{275a}$$

$$= \frac{3mk_BT}{(i\omega+\epsilon)+\xi}, \tag{275b}$$

where $z = i\omega+\epsilon$ lies in the right half of the complex plane.

Taking the limit $\epsilon \to 0^+$, we find that the real part of the spectral density $\mathcal{C}(i\omega) = \lim_{\epsilon\to 0^+}\mathcal{C}(i\omega+\epsilon)$ is a Lorentzian with a maximum at $\omega = 0$ and a half-width of 2ξ. The half-width provides a measure of the correlation time $\tau_C = \xi^{-1}$ for momentum fluctuations:

$$\tau_C = \left(\frac{1}{3mk_BT}\right)\lim_{z\to 0^+}\mathcal{C}_{pp}(z). \tag{276}$$

The result given by Eq. (276) and the relation $D = k_BT/m\xi$ enable us to write down a formal expression for the diffusion coefficient D:

$$D = \left(\frac{1}{3m^2}\right)\lim_{z\to 0^+}\int_0^\infty dt\,\exp(-zt)\langle\vec{p}(t)\cdot\vec{p}(0)\rangle_{\text{eq}}. \tag{277}$$

Spontaneous equilibrium fluctuations in the force $\vec{f}(t)$ are described by the equilibrium time correlation function

$$C_{ff}(t) = \langle\vec{f}(t)\cdot\vec{f}(0)\rangle_{\text{eq}} \tag{278a}$$

$$= 6mk_BT\xi\delta(t). \tag{278b}$$

Such fluctuations are δ-correlated in time by assumption.

The spectral density $\mathcal{C}_{ff}(i\omega+\epsilon)$ of the force fluctuations is given by

$$\mathcal{C}_{ff}(i\omega+\epsilon) = \int_0^\infty dt\,\exp[-(i\omega+\epsilon)t]\langle\vec{f}(t)\cdot\vec{f}(0)\rangle_{\text{eq}} \tag{279a}$$

$$= 6mk_BT\xi. \tag{279b}$$

Since the force $\vec{f}(t)$ is δ-correlated in time, the spectral density $\mathcal{C}_{ff}(i\omega+\epsilon)$ is completely flat, i.e., frequency independent. Such a spectrum is called a white spectrum, and its source is referred to as white noise.

It is clear from Eqs. (279a) and (279b) that the friction coefficient ξ is given by

$$\xi = \left(\frac{1}{6mk_BT}\right)\lim_{z\to 0^+}\int_0^\infty dt\,\exp(-zt)\langle\vec{f}(t)\cdot\vec{f}(0)\rangle_{\text{eq}}. \tag{280}$$

From this result, we see that the frictional damping of the particle's momentum arises from spontaneous equilibrium fluctuations in the force acting on the particle. This is sometimes called a fluctuation–dissipation relation.

Now that we have completed our exploration of Langevin dynamics, let us return to the use of Langevin

dynamics to determine the "drift" and "diffusion" coefficients appearing in the linear Fokker–Planck equation. Substituting Eqs. (270) and (271) into Eqs. (261) and (262), one can demonstrate that the following are the only nonnegligible "drift" and "diffusion" coefficients for time scales Δt much shorter than the time scale ξ^{-1} for momentum relaxation, i.e., $\xi \Delta t \ll 1$:

$$\mathbb{K}^{(1)}_{p_\alpha}(\vec{\Gamma}^N) = -\xi p_\alpha, \tag{281}$$

$$\mathbb{K}^{(1)}_{q_\alpha}(\vec{\Gamma}^N) = \frac{p_\alpha}{m}, \tag{282}$$

and

$$\mathbb{K}^{(1)}_{p_\alpha, p_\alpha}(\vec{\Gamma}^N) = 2mk_\mathrm{B}T\xi. \tag{283}$$

For this limit, the linear Fokker–Planck equation given by Eq. (258) assumes the form of the Fokker–Planck equation given by Eqs. (245)–(247) for the case of a single particle with an isotropic friction tensor and in the absence of a force field due to other particles undergoing Brownian motion.

If the time scale Δt is much longer than the time scale ξ^{-1} for momentum relaxation, all of the "drift" and "diffusion" coefficients are negligible except for

$$\mathbb{K}^{(2)}_{q_\alpha, q_\alpha}(\vec{\Gamma}^N) = \frac{2k_\mathrm{B}T}{m\xi}. \tag{284}$$

For this limit, the linear Fokker–Planck equation assumes the form of the Smoluchowski equation given by Eq. (248) for the case of a single particle with an isotropic diffusion tensor and in the absence of a force field due to other particles undergoing Brownian motion. One can also show that the diffusion coefficient $D = k_\mathrm{B}T/m\xi$ is given by

$$D = \lim_{\Delta t \to \infty} \left[\frac{\langle |\vec{q}(\Delta t) - \vec{q}(0)|^2 \rangle_B}{6\Delta t} \right]. \tag{285}$$

C. Nonequilibrium Thermodynamics (Phenomenological Theory of Irreversible Processes)

According to thermodynamics, the entropy change dS associated with an infinitesimal transformation of a system is given by

$$dS = d_\mathrm{i}S + d_\mathrm{e}S, \tag{286}$$

where $d_\mathrm{i}S$ is the entropy change inside the system and $d_\mathrm{e}S$ is the entropy transfer across the boundary between the system and its surroundings. The second law of thermodynamics states that $d_\mathrm{i}S$ is positive for irreversible transformations and zero for reversible transformations:

$$d_\mathrm{i}S \geq 0. \tag{287}$$

The entropy transfer $d_\mathrm{e}S$ may be positive, negative, or zero.

For an adiabatically insulated system (a system that does not undergo any transfer processes with its surroundings), $d_\mathrm{e}S = 0$. Then

$$dS \geq 0. \tag{288}$$

This is a well-known form of the second law of thermodynamics.

For a closed system that can only exchange heat with its surroundings,

$$d_\mathrm{e}S = \frac{dQ}{T}, \tag{289}$$

where dQ is the heat received by the system at the absolute temperature T. Then

$$dS \geq \frac{dQ}{T}. \tag{290}$$

This is another well-known version of the second law of thermodynamics. In general, Eqs. (286) and (287) imply $dS \geq d_\mathrm{e}S$.

According to the second law of thermodynamics, any spontaneous process in an isolated system out of equilibrium will lead to an increase in the entropy inside that system. In spite of the general validity of the second law, we have yet to fully understand the problem of irreversible time evolution. Of course, Brownian motion theory does provide some insight into the direction that might be pursued in seeking answers to this problem. The most important idea to emerge from Brownian motion theory is the notion that dissipative or irreversible behavior arises from spontaneous equilibrium fluctuations.

Although thermodynamics allows us to make statements about the entropy change associated with the passage of a system from one constrained equilibrium state to another, it lacks the appropriate mathematical apparatus for treating the irreversible time evolution of systems out of equilibrium. Thus, the incorporation of irreversible processes into thermodynamics is incomplete. In part, this problem stems from the lack of any explicit reference to time in the formalism of thermodynamics.

There have been many attempts to incorporate irreversible processes into thermodynamics. Early attempts were confined to the treatment of special classes of phenomena, such as thermoelectric effects. In these attempts, the theoretical formulation of a given irreversible process was usually done in a manner that was disjoint from the formulation of other irreversible processes, each formulation requiring different ad hoc assumptions about the nature of a system. This lack of unity in the theory of irreversible processes existed until Onsager presented a unifying conceptual framework for linear Markovian (without

memory) systems. The work of Onsager and later work done in the same spirit by Prigogine, de Groot, Mazur, and others is usually called the phenomenological theory of irreversible processes or nonequilibrium thermodynamics. Henceforth, we shall use these labels interchangeably.

In nonequilibrium thermodynamics, one usually divides systems into two distinct types, namely, discrete composite systems and continuous systems. Discrete composite systems are systems for which the time-dependent thermodynamic properties are discontinuous across the boundaries between the subsystems constituting the composite system. Continuous systems are systems for which thermodynamic properties vary continuously throughout.

In this section, we first discuss the basic results from nonequilibrium thermodynamics for continuous and discrete systems. Then we show that all of these results emerge from a quantum statistical mechanical treatment of nonequilibrium systems based on a time-dependent version of the maximum entropy principle. It is shown that this treatment not only includes nonequilibrium thermodynamics, but also provides a more general framework for discussing both equilibrium and nonequilibrium systems.

1. Continuous Systems

Consider a continuous, chemically nonreactive system characterized by a set of local densities $\{\rho_j(\vec{r},t)\}$ associated with conserved quantities. For such a system, the relevant local densities might include particle density, momentum density, and energy density. The time evolution of the local densities is governed by the continuity equation

$$\frac{\partial \rho_j(\vec{r},t)}{\partial t} = -\vec{\nabla}\cdot\vec{J}_j(\vec{r},t), \tag{291}$$

where the current density $\vec{J}_j(\vec{r},t)$ describes the flow associated with the local density $\rho_j(\vec{r},t)$ at the space-time point $(\vec{r},t)$.

In nonequilibrium thermodynamics, the entropy production $P_S(t) = dS(t)/dt$ associated with the evolution of the system is given by

$$P_S(t) = P_S^{(\mathrm{i})}(t) + P_S^{(\mathrm{e})}(t). \tag{292}$$

In the above, $P_S^{(\mathrm{i})}(t) = d_\mathrm{i}S(t)/dt$ is the entropy production due to irreversible processes inside the system and $P_S^{(\mathrm{e})}(t) = d_\mathrm{e}S(t)/dt$ is the entropy production due to entropy flow across the boundary between the system and its surroundings.

The contributions to Eq. (292) are given by

$$P_S(t) = \int_{\mathcal{D}} dV \frac{\partial \rho_s(\vec{r},t)}{\partial t}, \tag{293}$$

$$P_S^{(\mathrm{i})}(t) = \int_{\mathcal{D}} dV \sigma_s(\vec{r},t) \tag{294a}$$

$$= \int_{\mathcal{D}} dV \frac{d\rho_s(\vec{r},t)}{dt}, \tag{294b}$$

and

$$P_S^{(\mathrm{e})}(t) = -\int_{\mathcal{S}} dS\hat{n}\cdot\vec{J}_s(\vec{r},t). \tag{295}$$

In the above, $\rho_s(\vec{r},t)$ is the entropy density, $\sigma_s(\vec{r},t)$ the entropy source strength, and $\vec{J}_s(\vec{r},t)$ the entropy current density. The volume integrations in Eqs. (293) (294b) are over the spatial domain $\mathcal{D}$ occupied by the system. The surface integration in Eq. (295) is over the boundary surface $\mathcal{S}$ separating the system from its surroundings. The unit vector $\hat{n}$ is normal to $\mathcal{S}$ and directed outward.

The time evolution of the entropy density $\rho_s(\vec{r},t)$ is assumed to be governed by the entropy balance equation

$$\frac{d\rho_s(\vec{r},t)}{dt} = \frac{\partial \rho_s(\vec{r},t)}{\partial t} + \vec{\nabla}\cdot\vec{J}_s(\vec{r},t). \tag{296}$$

The total time derivative $d\rho_s(\vec{r},t)/dt$ is the rate at which the entropy density is being produced. The quantities $\partial\rho_s(\vec{r},t)/\partial t$ and $\vec{\nabla}\cdot\vec{J}_s(\vec{r},t)$ represent the rates at which the entropy density increases and decreases, respectively, in the neighborhood of $\vec{r}$.

Since the thermodynamic properties vary throughout a continuous system, the following Gibbsian relation is assumed:

$$d\rho_s(\vec{r},t) = k_\mathrm{B}\sum_j \Lambda_j(\vec{r},t)d\rho_j(\vec{r},t), \tag{297}$$

where $k_\mathrm{B}\Lambda_j(\vec{r},t)$ is regarded as a local thermodynamic parameter. This relation allows one to identify the local thermodynamic parameters with the partial derivatives of the entropy density, i.e.,

$$k_\mathrm{B}\Lambda_j(\vec{r},t) = \frac{\partial \rho_s(\vec{r},t)}{\partial \rho_j(\vec{r},t)}. \tag{298}$$

Adopting the above-described picture of continuous systems, one can demonstrate that

$$\frac{\partial \rho_s(\vec{r},t)}{\partial t} = k_\mathrm{B}\sum_j \Lambda_j(\vec{r},t)\frac{\partial \rho_j(\vec{r},t)}{\partial t}, \tag{299}$$

$$\sigma_s(\vec{r},t) = \frac{d\rho_s(\vec{r},t)}{dt} \tag{300a}$$

$$= \sum_j \vec{X}_j(\vec{r},t)\cdot\vec{J}_j(\vec{r},t), \tag{300b}$$

and

$$\vec{J}_s(\vec{r},t) = k_\mathrm{B}\sum_j \Lambda_j(\vec{r},t)\vec{J}_j(\vec{r},t), \tag{301}$$

where

$$\vec{X}_j(\vec{r},t) = k_B \vec{\nabla} \Lambda_j(\vec{r},t). \tag{302}$$

In nonequilibrium thermodynamics, the quantities $\{\vec{X}_j(\vec{r},t)\}$ are viewed as thermodynamic driving forces that give rise to the flows described by the current densities $\{\vec{J}_j(\vec{r},t)\}$. For systems in equilibrium, the thermodynamic parameters $\{k_B\Lambda_j(\vec{r},t)\}$ are spatially uniform. Then the driving forces $\{X_j(\vec{r},t)\}$ vanish. Consequently, the current densities $\{\vec{J}_j(\vec{r},t)\}$ vanish for systems in equilibrium.

Making use of Eqs. (293)–(295) and (299)–(302), one finds that the contributions to Eq. (292) are given by

$$P_S(t) = k_B \sum_j \int_{\mathcal{D}} dV \Lambda_j(\vec{r},t) \frac{\partial \rho_j(\vec{r},t)}{\partial t}, \tag{303}$$

$$P_S^{(i)}(t) = \sum_j \int_{\mathcal{D}} dV \vec{X}_j(\vec{r},t) \cdot \vec{J}_j(\vec{r},t), \tag{304}$$

and

$$P_S^{(e)}(t) = -k_B \sum_j \int_{\mathcal{S}} dS \Lambda_j(\vec{r},t) \hat{n} \cdot \vec{J}_j(\vec{r},t). \tag{305}$$

From the above, we find that the entropy production inside the system, $P_S^{(i)}(t)$, is due to the internal flows described by $\{\vec{J}_j(\vec{r},t)\}$. Also, we find that the contribution $P_S^{(e)}(t)$ is due to the flows across the boundary between the system and its surroundings.

In view of Eq. (303), the entropy production $P_S(t)$ must vanish when the system is in a stationary state, i.e., a state for which the local densities $\rho_j(\vec{r},t)$ are time independent and the current densitites $\vec{J}_j(\vec{r},t)$ are spatially uniform.

If the system is in an equilibrium stationary state (state of equilibrium), we find from Eq. (304) that the entropy production inside the system, $P_S^{(i)}(t)$, vanishes. Since the total entropy production $P_S(t)$ must also vanish, we find from Eq. (292) that there is no entropy transfer across the boundary between the system and its surroundings when the system is in an equilibrium stationary state.

For nonequilibrium stationary states, the entropy production inside the system, $P_S^{(i)}(t)$, is nonvanishing. Since the total entropy production $P_S(t)$ must vanish for a nonequilibrium stationary state, we find from Eqs. (292), (304), and (305) that the entropy production inside the system is maintained by the entropy flow across the boundary between the system and its surroundings. This entropy flow can be regarded as maintaining the nonequilibrium stationary state.

It is instructive to rewrite Eq. (292) as

$$dS(t) = d_i S(t) + d_e S(t), \tag{306}$$

where

$$d_i S(t) = \sum_j \int_{\mathcal{D}} dV \vec{X}_j(\vec{r},t) \cdot \vec{J}_j(\vec{r},t)\, dt \tag{307}$$

and

$$d_e S(t) = -k_B \sum_j \int_{\mathcal{S}} dS \Lambda_j(\vec{r},t)\, \hat{n} \cdot \vec{J}_j(\vec{r},t)\, dt. \tag{308}$$

In the above, $d_i S(t)$ is the entropy change due to the flows inside the system in an infinitesimal time interval dt centered about the time t. During this time interval, the entropy change arising from the flows across the boundary between the system and its surroundings is given by $d_e S(t)$.

For a closed system that can only exchange energy with its surroundings, we can write

$$d_e S(t) = -\int_{\mathcal{S}} dS \left[\frac{\hat{n} \cdot \vec{J}_E(\vec{r},t)}{T(\vec{r},t)} \right] dt, \tag{309}$$

where $\vec{J}_E(\vec{r},t)$ is the energy current density (heat flux) and $T(\vec{r},t) = 1/[k_B\Lambda_E(\vec{r},t)]$ is a spatially dependent temperature, with $k_B\Lambda_E(\vec{r},t)$ denoting the thermodynamic parameter conjugate to the energy density $\rho_E(\vec{r},t)$.

If the temperature $T(\vec{r},t)$ on the boundary surface $\mathcal{S}$ is spatially uniform and held fixed at T, the above expression for $d_e S(t)$ assumes the usual form

$$d_e S(t) = \frac{dQ(t)}{T} \tag{310}$$

given in thermodynamics for the entropy change associated with the heat

$$dQ(t) = -\int_{\mathcal{S}} dS \hat{n} \cdot \vec{J}_E(\vec{r},t)\, dt \tag{311}$$

received by a closed system from its surroundings at the absolute temperature T.

Making use of Eqs. (292), (304), and (305), we write the rate of entropy production as

$$\frac{dP_S(t)}{dt} = \frac{dP_S^{(i)}(t)}{dt} + \frac{dP_S^{(e)}(t)}{dt}, \tag{312}$$

where

$$\frac{dP_S^{(i)}(t)}{dt} = \frac{d_X P_S^{(i)}(t)}{dt} + \frac{d_J P_S^{(i)}(t)}{dt} \tag{313}$$

and

$$\frac{dP_S^{(e)}(t)}{dt} = \frac{d_\Lambda P_S^{(e)}(t)}{dt} + \frac{d_J P_S^{(e)}(t)}{dt}, \tag{314}$$

with

$$\frac{d_X P_S^{(i)}(t)}{dt} = \sum_j \int_{\mathcal{D}} dV \vec{J}_j(\vec{r},t) \cdot \frac{\partial \vec{X}_j(\vec{r},t)}{\partial t}, \tag{315}$$

$$\frac{d_J P_S^{(\mathrm{i})}(t)}{dt} = \sum_j \int_{\mathcal{D}} dV \frac{\partial \vec{J}_j(\vec{r},t)}{dt} \cdot \vec{X}_j(\vec{r},t), \quad (316)$$

$$\frac{d_\Lambda P_S^{(\mathrm{e})}(t)}{dt} = -k_\mathrm{B} \sum_j \int_{\mathcal{S}} dS\hat{n} \cdot \vec{J}_j(\vec{r},t) \frac{\partial \Lambda_j(\vec{r},t)}{\partial t}, \quad (317)$$

and

$$\frac{d_J P_S^{(\mathrm{e})}(t)}{dt} = -k_\mathrm{B} \sum_j \int_{\mathcal{S}} dS\hat{n} \cdot \frac{\partial \vec{J}_j(\vec{r},t)}{\partial t} \Lambda_j(\vec{r},t). \quad (318)$$

The quantity $[d_X P_S^{(\mathrm{i})}(t)/dt][d_J P_S^{(\mathrm{i})}(t)/dt]$ is the rate of entropy production inside the system due to the driving forces $\vec{X}_j(\vec{r},t)$ [current densities $\vec{J}_j(\vec{r},t)$] that change with time. Similarly, $[d_\Lambda P_S^{(\mathrm{e})}(t)/dt][d_J P_S^{(\mathrm{e})}(t)/dt]$ is the rate of entropy production due to the thermodynamic parameters $k_\mathrm{B}\Lambda_j(\vec{r},t)$ [current densities $\vec{J}_j(\vec{r},t)$] that change with time on the boundary surface $\mathcal{S}$ between the system and its surroundings.

In nonequilibrium thermodynamics, it is assumed that the rate of entropy production inside the system due to the time variation of the driving forces $\vec{X}_j(\vec{r},t)$ decreases in the course of time, i.e.,

$$\frac{d_X P_S^{(\mathrm{i})}(t)}{dt} \le 0. \quad (319)$$

We shall refer to this inequality, first established by Glansdorff and Prigogine, as the Glansdorff–Prigogine evolution theorem.

In the description of transport processes in homogenous systems, investigators often adopt linear phenomenological equations of the form

$$\vec{J}_j(\vec{r},t) = \sum_k L_{jk} \vec{X}_k(\vec{r},t). \quad (320)$$

Phenomenological equations possessing this form include Fourier's law of heat conduction, Fick's law of diffusion, and Ohm's law. Systems characterized by such equations are said to be linear systems.

It is usually assumed that the phenomenological coefficients $\{L_{jk}\}$ satisfy Onsager's reciprocity relation

$$L_{jk} = L_{kj}. \quad (321)$$

In addition, it is assumed that the matrix $\mathbf{L}$ formed by the phenomenological coefficients is positive definite, i.e., $\sum_{j,k} \Phi_j L_{jk} \Phi_k > 0$ for any set of arbitrary scalar quantities $\{\Phi_j\}$.

Adopting the above-described model for linear systems, one can demonstrate that

$$P_S^{(\mathrm{i})}(t) \ge 0 \quad (322)$$

and

$$\frac{d_X P_S^{(\mathrm{i})}(t)}{dt} = \frac{d_J P_S^{(\mathrm{i})}(t)}{dt} = \frac{1}{2}\frac{d P_S^{(\mathrm{i})}(t)}{dt} \le 0. \quad (323)$$

The first inequality follows from the linear phenomenological equations and the assumed character of the phenomenological coefficients. The second inequality follows from the Glansdorff–Prigogine evolution theorem.

According to Eq. (323), the enropy production $P_S^{(\mathrm{i})}(t)$ must decrease in time and assume a minimum value when a stationary state is reached. This result, called the theorem of minimum entropy production, was established by Glansdorff and Progogine.

In general, we can say nothing about the sign of $d_J P_S^{(\mathrm{i})}(t)/dt$ and hence the sign of $d P_S^{(\mathrm{i})}(t)/dt$ for nonlinear systems. Thus, stationary states for nonlinear systems do not necessarily represent states of minimum entropy production.

2. Discrete Composite Systems

Consider an adiabatically isolated composite system made up of two subsystems. The two subsystems, designated by $\alpha = 1$ and 2, are characterized by the same set of thermodynamic coordinates $\{O_j^{(\alpha)}(t)\}$. These thermodynamic coordinates are assumed to represent conserved quantities such as energy and particle number. Hence, they conform to the conservation relations

$$O_j = O_j^{(1)}(t) + O_j^{(2)}(t), \quad (324)$$

where O_j is time independent.

In nonequilibrium thermodynamics, the entropy production $P_S(t)$ associated with transfer processes between subsystems 1 and 2 is written

$$P_S(t) = \sum_j J_j(t) X_j(t), \quad (325)$$

where

$$J_j(t) = \frac{d O_j^{(1)}(t)}{dt} = -\frac{d O_j^{(2)}(t)}{dt} \quad (326)$$

and

$$X_j(t) = k_\mathrm{B}\left[\Lambda_j^{(1)}(t) - \Lambda_j^{(2)}(t)\right], \quad (327)$$

with $k_\mathrm{B}\Lambda_j^{(\alpha)}(t)$ denoting the thermodynamic parameter conjugate to the thermodynamic coordinate $O_j^{(\alpha)}(t)$.

The quantities $\{X_j(t)\}$, called thermodynamic driving forces, are said to give rise to the flows $\{J_j(t)\}$ between subsystems 1 and 2. For composite systems in equilibrium, $\Lambda_j^{(1)}(t) = \Lambda_j^{(2)}(t)$. Then the driving forces $\{X_j(t)\}$ vanish. Consequently, the flows $\{J_j(t)\}$ vanish for composite systems in equilibrium.

The thermodynamic driving forces are assumed to be given by

$$X_j(t) = \frac{\partial S(t)}{\partial O_j^{(1)}(t)}. \tag{328}$$

This implies

$$dS(t) = \sum_j X_j(t)\, dO_j^{(1)}(t). \tag{329}$$

It follows from Eq. (325) that the rate of entropy production can be written

$$\frac{dP_S(t)}{dt} = \frac{d_X P_S(t)}{dt} + \frac{d_J P_S(t)}{dt}, \tag{330}$$

where

$$\frac{d_X P_S(t)}{dt} = \sum_j J_j(t) \frac{dX_j(t)}{dt} \tag{331}$$

and

$$\frac{d_J P_S(t)}{dt} = \sum_j \frac{dJ_j(t)}{dt} X_j(t). \tag{332}$$

The contribution to the entropy production due to the time variation of the driving forces $\{X_j(t)\}$ is assumed to conform to the following version of the Glansdorff–Prigogine evolution theorem:

$$\frac{d_X P_S(t)}{dt} \leq 0. \tag{333}$$

In the description of transfer processes between the two subsystems, investigators often adopt linear phenomenological equations of the form

$$J_j(t) = \sum_k L_{jk} X_k(t). \tag{334}$$

Phenomenological equations possessing this form have been used to describe thermo osmosis and thermoelectric phenomena.

As for continuous systems, the following assumptions are made: (i) The phenomenological coefficients satisfy Onsager's reciprocity relation $L_{jk} = L_{kj}$. (ii) The matrix **L** formed by the phenomenological coefficients is positive definite.

For composite systems conforming to the above-described requirements, one can demonstrate that

$$P_S(t) \geq 0 \tag{335}$$

and

$$\frac{d_X P_S(t)}{dt} = \frac{d_J P_S(t)}{dt} = \frac{1}{2}\frac{dP_S(t)}{dt} \leq 0. \tag{336}$$

The latter relation implies that linear composite systems satisfy the theorem of minimum entropy production.

3. Onsager's Linear Phenomenological Theory

As indicated earlier, linear nonequilibrium thermodynamics is based on the following postulates: (i) A Gibbsian expression for the entropy change $dS(t)$ is valid for systems out of equilibrium. (ii) The entropy production is given by a bilinear form in the drives forces and flows. (iii) The flows can be expressed as a linear combination of the driving forces. (iv) The phenomenological coefficients satisfy Onsager's reciprocity relations.

In the proof of the reciprocity relations $L_{jk} = L_{kj}$, Onsager wrote the linear phenomenological equations for the case of observables with a discrete index in the form

$$d\overline{O_j(t + \Delta t \mid \mathbf{O}')}/dt = \sum_k L_{jk} \partial S(\mathbf{O}')/\partial O_k', \tag{337}$$

where $d\overline{O_j(t + \Delta t \mid \mathbf{O}')}/dt$ is the phenomenological time derivative

$$d\overline{O_j(t + \Delta t \mid \mathbf{O}')}/dt = [\overline{O_j(t + \Delta t \mid \mathbf{O}')} - \overline{O_j(t \mid \mathbf{O}')}]/\Delta t, \tag{338}$$

with $O_j(t + \Delta t \mid \mathbf{O}')$ representing the average value of the observable O_j at time $t + \Delta t$ given that the set of observables **O** possess the values $\mathbf{O}'$ at time t.

To proceed further, Onsager did not take the Gibbsian path. Instead, he adopted Boltzmann's definition of entropy and Einstein's theory of fluctuations. Nonetheless, Onsager was led to the following expression for the phenomenological coefficient L_{jk}:

$$L_{jk} = -\left(\frac{1}{k_B \Delta t}\right)[\overline{O_j(t + \Delta t) O_k(t)} - \overline{O_j(t) O_k(t)}], \tag{339}$$

where the quantity $\overline{O_j(t + \Delta t) O_k(t)}$ is an averaged quantity intended to represent the correlation between the events $O_j(t + \Delta t)$ and $O_k(t)$. In defining the average $\overline{O_j(t + \Delta t) O_k(t)}$, Onsager made an assumption reminiscent of a basic assumption made in Brownian motion theory. More specifically, Onsager assumed that stochastic averaging is equivalent to ensemble averaging. Making use of this assumption and arguments based on the time-reversal invariance of the microscopic equations of motion, Onsager concluded that Eq. (339) implies $L_{jk} = L_{kj}$.

D. Statistical Mechanical Basis for Nonequilibrium Thermodynamics

Consider some arbitrary quantum mechanical system out of equilibrium. Assuming the validity of the Gibbsian definition of entropy, we obtain for the entropy $S(t)$ of the system at time t

$$S(t) = -k_B \mathrm{Tr}\, \hat{\bar{\rho}}(t) \ln\, \hat{\bar{\rho}}(t), \tag{340}$$

where the statistical density operator $\hat{\bar{\rho}}(t)$ describes the macroscopic state of the system.

The macroscopic state of the system is defined by specifying the values of certain thermodynamic coordinates. The thermodynamic coordinates may or may not depend on position and/or time. For the sake of simplicity, we confine our attention to systems that are characterized by either spatially independent thermodynamic coordinates $\{\langle \hat{O}_j(t)\rangle\}$ or spatially dependent thermodynamic coordinates $\{\langle \hat{O}_j(\vec{r},t)\rangle\}$. In general, both types of thermodynamic coordinates are required to specify the macroscopic state of a system.

Independent of the nature of the thermodynamic coordinates, we assume that the maximum entropy principle can be used to construct the statistical density operator $\hat{\bar{\rho}}(t)$. Thus, $\hat{\bar{\rho}}(t)$ is the maximum entropy density operator consistent with the information defining the macroscopic state of the system. As for equilibrium systems, we require the thermodynamic coordinates to be linearly independent.

1. Spatially Independent Thermodynamic Coordinates

Making use of the time-dependent version of the maximum entropy principle, we find that the statistical density operator $\hat{\bar{\rho}}(t)$ for a system characterized by the spatially independent thermodynamic coordinates $\{\langle \hat{O}_j(t)\rangle\}$ is given by

$$\hat{\bar{\rho}}(t) = \exp\left[\Omega(t)\hat{I} - \sum_j \Lambda_j(t)\hat{O}_j\right], \tag{341}$$

where

$$\Omega(t) = -\ln Z(t), \tag{342}$$

with

$$Z(t) = \mathrm{Tr}\exp\left[-\sum_j \Lambda_j(t)\hat{O}_j\right]. \tag{343}$$

Making use of Eq. (340), we write the entropy $S(t)$ of the macroscopic state described by $\hat{\bar{\rho}}(t)$ as

$$S(t) = -k_B\Omega(t) + k_B\sum_j \Lambda_j(t)\langle \hat{O}_j(t)\rangle. \tag{344}$$

In the spirit of the statistical mechanical treatment of equilibrium systems, we interpret $\Omega(t)$ and $\Phi(t) = -k_B\Omega(t)$ as thermodynamic potentials and $\{\Lambda_j(t)\}$ as thermodynamic parameters.

Similar to equilibrium systems, the macroscopic state of the system can be defined in terms of either the thermodynamic parameters $\{\Lambda_j(t)\}$ or the thermodynamic coordinates $\{\langle \hat{O}_j(t)\rangle\}$. Both sets of variables represent the same information. If the thermodynamic parameters are known, the set of equations

$$\frac{\partial S(t)}{\partial\langle \hat{O}_j(t)\rangle} = k_B\Lambda_j(t) \tag{345}$$

can be solved for the averages. If the averages are known, the set of equations

$$\frac{\partial\Phi(t)}{\partial\Lambda_j(t)} = -k_B\langle \hat{O}_j(t)\rangle \tag{346}$$

can be solved for the thermodynamic parameters.

Adopting a formal development similar to the one given in the statistical mechanical treatment of equilibrium systems, one finds that the basic thermodynamic relations for nonequilibrium systems are identical in form to the basic thermodynamic relations for equilibrium systems, except that all of the thermodynamic quantities are time dependent. For example, the entropy $S(t)$ and the thermodynamic potential $\Phi(t)$ for nonequilibrium systems are connected by the generalized Gibbs–Helmholz relations

$$S(t) = \Phi(t) - \sum_j \Lambda_j(t)\frac{\partial\Phi(t)}{\partial\Lambda_j(t)} \tag{347}$$

and

$$\Phi(t) = S(t) - \sum_j \langle \hat{O}_j(t)\rangle\frac{\partial S(t)}{\partial\langle \hat{O}_j(t)\rangle}. \tag{348}$$

For the case of nonequilibrium systems, the generalized Maxwell relations assume the forms

$$\frac{\partial\Lambda_k(t)}{\partial\langle \hat{O}_j(t)\rangle} = \frac{\partial\Lambda_j(t)}{\partial\langle \hat{O}_k(t)\rangle} \tag{349}$$

and

$$\frac{\partial\langle \hat{O}_k(t)\rangle}{\partial\Lambda_j(t)} = \frac{\partial\langle \hat{O}_j(t)\rangle}{\partial\Lambda_k(t)}. \tag{350}$$

In addition, we can write the differentials $dS(t)$, $d\Omega(t)$, and $d\Phi(t)$ as

$$dS(t) = k_B\sum_j \Lambda_j(t)\, d\langle \hat{O}_j(t)\rangle, \tag{351}$$

$$d\Omega(t) = \sum_j \langle \hat{O}_j(t)\rangle\, d\Lambda_j(t), \tag{352}$$

and

$$d\Phi(t) = -k_B\sum_j \langle \hat{O}_j(t)\rangle\, d\Lambda_j(t). \tag{353}$$

For the case of discrete composite systems, the result given by Eq. (351) is equivalent to the Gibbsian form assumed for the differential $dS(t)$ of the entropy $S(t)$ in nonequilibrium thermodynamics. [See Eq. (329).] It follows that the thermodynamic driving forces $\{X_j(t)\}$ in nonequilibrium thermodynamics are given by Eq. (328).

The time dependence of the entropy $S(t)$ is due to its dependence on thermodynamic coordinates $\{\langle \hat{O}_j(t)\rangle\}$. Making use of this observation and Eq. (345), we find that the entropy production $P_S(t)$ is given by

$$P_S(t) = k_B \sum_j \Lambda_j(t) \frac{d\langle \hat{O}_j(t)\rangle}{dt}. \tag{354}$$

For the case of discrete composite systems, this equation is equivalent to the bilinear form assumed for the entropy production in nonequilibrium thermodynamics. [See Eq. (325).]

Making use of Eq. (354), we write the rate of entropy production as

$$\frac{dP_S(t)}{dt} = \frac{d_\Lambda P_S(t)}{dt} + \frac{d_{\dot{o}} P_S(t)}{dt}, \tag{355}$$

where

$$\frac{d_\Lambda P_S(t)}{dt} = k_B \sum_j \left[\frac{d\Lambda_j(t)}{dt}\right]\left[\frac{d\langle \hat{O}_j(t)\rangle}{dt}\right] \tag{356}$$

and

$$\frac{d_{\dot{o}} P_S(t)}{dt} = k_B \sum_j \Lambda_j(t)\left[\frac{d^2\langle \hat{O}_j(t)\rangle}{dt^2}\right]. \tag{357}$$

One can show that

$$\frac{\partial\langle \hat{O}_j(t)\rangle}{\partial \Lambda_k(t)} = -\chi_{jk}(t) \tag{358}$$

and

$$\frac{\partial \Lambda_j(t)}{\partial\langle \hat{O}_k(t)\rangle} = -\chi^{-1}_{jk}(t). \tag{359}$$

In the above, $\chi^{-1}_{jk}(t)$ denotes a matrix element of the inverse $\boldsymbol{\chi}^{-1}(t)$ of the matrix $\boldsymbol{\chi}(t)$. The matrix elements $\chi_{jk}(t)$ of $\boldsymbol{\chi}(t)$ are given by

$$\chi_{jk}(t) = \langle \delta_t \hat{O}_k \delta_t \hat{\hat{O}}_j\rangle_{\hat{\rho}(t)} \tag{360a}$$

$$= \mathrm{Tr}\ \hat{\rho}(t)\delta_t \hat{O}_k \delta_t \hat{\hat{O}}_j, \tag{360b}$$

where

$$\delta_t \hat{\hat{O}}_j = \hat{\hat{O}}_j - \langle \hat{\hat{O}}_j(t)\rangle \tag{361}$$

and

$$\delta_t \hat{O}_k = \hat{O}_k - \langle \hat{O}_k(t)\rangle, \tag{362}$$

with

$$\hat{\hat{O}}_j = \int_0^1 d\lambda \hat{\rho}(t)^\lambda \hat{O}_j \hat{\rho}(t)^{-\lambda}. \tag{363}$$

In general, $\langle \hat{\hat{O}}_j(t)\rangle = \langle \hat{O}_j(t)\rangle$.

The matrix elements $\chi_{jk}(t)$ of $\boldsymbol{\chi}(t)$ can be regarded as generalized nonequilibrium susceptibilities. For systems in thermal equilibrium, $\hat{\hat{\rho}}(t)$ becomes the equilibrium canonical density operator. Thus, the susceptibilities $\chi_{jk}(t)$ become equivalent to the static susceptibilities of equilibrium statistical mechanics when the system is in thermal equilibrium. The susceptibilities $\chi_{jk}(t)$ describe the correlation between the fluctuations in the thermodynamic coordinates for nonequilibrium systems.

Making use of Eqs. (360a)–(363), one can demonstrate that $\boldsymbol{\chi}(t)$ is a real, symmetric, nonnegative-definite matrix. By nonnegative-definite matrix, we mean that $\sum_j \Phi^*_j(t)\chi_{jk}(t)\Phi_k(t) \geq 0$ for any set of complex scalar quantities $\{\Phi_j(t)\}$ or $\sum_j \phi_j(t)\chi_{jk}(t)\phi_k(t) \geq 0$ for any set of real scalar quantities $\{\phi_j(t)\}$.

Since $\boldsymbol{\chi}(t)$ is a real, symmetric, nonnegative-definite matrix, its eigenvalues are real and nonnegative. If $\boldsymbol{\chi}(t)$ possesses a zero eigenvalue, it is singular and, consequently, the inverse of $\boldsymbol{\chi}^{-1}(t)$ of $\boldsymbol{\chi}(t)$ does not exist. Assuming that $\boldsymbol{\chi}(t)$ does not possess a zero eigenvalue, we conclude that $\boldsymbol{\chi}(t)$ and $\boldsymbol{\chi}^{-1}(t)$ are positive definite.

Of course, the matrix elements $\chi^{-1}_{jk}(t)$ of $\boldsymbol{\chi}^{-1}(t)$ can become very small when the fluctuations, as described by the matrix elements $\chi_{jk}(t)$ of $\boldsymbol{\chi}(t)$, become very large. For such cases, the partial derivatives given by Eq. (359) can become very small. Below, we discuss the ramifications of this behavior in some detail.

Making use of Eqs. (358) and (359), one can demonstrate that

$$\frac{d\langle \hat{O}_j(t)\rangle}{dt} = -\sum_k \chi_{jk}(t)\left[\frac{d\Lambda_k(t)}{dt}\right] \tag{364}$$

and

$$\frac{d\Lambda_j(t)}{dt} = -\sum_k \chi^{-1}_{jk}(t)\left[\frac{d\langle \hat{O}_k(t)\rangle}{dt}\right]. \tag{365}$$

With these equations at our disposal, we can rewrite Eq. (356) as

$$\frac{d_\Lambda P_S(t)}{dt} = -k_B \sum_{j,k}\left[\frac{d\langle \hat{O}_j(t)\rangle}{dt}\right]\chi^{-1}_{jk}(t)\left[\frac{d\langle \hat{O}_k(t)\rangle}{dt}\right] \tag{366a}$$

$$= -k_B \sum_{j,k}\left[\frac{d\Lambda_j(t)}{dt}\right]\chi_{jk}(t)\left[\frac{d\Lambda_k(t)}{dt}\right]. \tag{366b}$$

Since $\boldsymbol{\chi}(t)$ and $\boldsymbol{\chi}^{-1}(t)$ are positive definite, these relations imply

$$\frac{d_\Lambda P_S(t)}{dt} \leq 0. \tag{367}$$

For the case of discrete composite systems, this inequality becomes equivalent to the inequality embodying the Prigogine–Glansdorff evolution theorem in nonequilibrium thermodynamics. [See Eq. (333).] The above generalization of the Prigogine–Glansdorff evolution theorem applies independent of the nature of the physical system.

Thus far, we have established that the following results from nonequilibrium thermodynamics emerge from the statistical mechanical treatment of nonequilibrium systems when the system can be regarded as a discrete composite system: (i) The entropy change $dS(t)$ arising from transfer processes is given by $dS(t) = \sum_j X_j(t)\, d\langle \hat{O}_j^{(1)}(t)\rangle$. Consequently, the driving forces can be identified as $X_j(t) = \partial S(t)/\partial\langle \hat{O}_j^{(1)}(t)\rangle$. (ii) The entropy production $P_S(t)$ is given by the bilinear form $P_S(t) = \sum_j X_j(t) J_j(t)$. (iii) The evolution of the system is such that the inequality $d_X P_S(t)/dt \le 0$ is satisfied.

Let us further explore the consequences of the statistical mechanical treatment of discrete composite systems. For such systems, it is convenient to rewrite Eq. (364) as

$$\begin{bmatrix} \dfrac{d\langle \hat{\mathbf{O}}^{(1)}(t)\rangle}{dt} \\ \dfrac{d\langle \hat{\mathbf{O}}^{(2)}(t)\rangle}{dt} \end{bmatrix} = -\begin{bmatrix} \boldsymbol{\chi}^{(11)}(t) & \boldsymbol{\chi}^{(12)}(t) \\ \boldsymbol{\chi}^{(21)}(t) & \boldsymbol{\chi}^{(22)}(t) \end{bmatrix} \begin{bmatrix} \dfrac{d\boldsymbol{\Lambda}^{(1)}(t)}{dt} \\ \dfrac{d\boldsymbol{\Lambda}^{(2)}(t)}{dt} \end{bmatrix}. \tag{368}$$

In the above, we have partitioned the susceptibilty matrix $\boldsymbol{\chi}(t)$ into four blocks. The elements of block $\boldsymbol{\chi}^{(\alpha\alpha')}(t)$ are given by

$$\chi_{jk}^{(\alpha\alpha')} = \left\langle \delta_t \hat{O}_k^{(\alpha')} \delta_t \hat{\hat{O}}_j{}^{(\alpha)} \right\rangle_{\hat{\rho}(t)}. \tag{369}$$

By construction, $\hat{\rho}(t) = \hat{\rho}^{(1)}(t)\hat{\rho}^{(2)}(t)$, $[\hat{\rho}^{(1)}(t), \hat{\rho}^{(2)}(t)]_- = \hat{0}$, and $[\hat{O}_j^{(1)}, \hat{O}_j^{(2)}]_- = \hat{0}$, where $\hat{0}$ is the null operator. Making use of these relations and Eq. (369), one finds that $\chi_{jk}^{(12)}(t) = \chi_{jk}^{(21)}(t) = 0$. This result expresses the lack of correlation between fluctuations in subsystems 1 and 2. Obviously, the blocks $\boldsymbol{\chi}^{(12)}(t)$ and $\boldsymbol{\chi}^{(21)}(t)$ of the susceptiblity matrix $\boldsymbol{\chi}(t)$ are null.

One can demonstrate that the nonnull blocks $\boldsymbol{\chi}^{(11)}(t)$ and $\boldsymbol{\chi}^{(22)}(t)$ of $\boldsymbol{\chi}(t)$ are endowed with the same character as $\boldsymbol{\chi}(t)$. More specifically, $\boldsymbol{\chi}^{(11)}(t)$ and $\boldsymbol{\chi}^{(22)}(t)$ are real, symmetric, positive-definite matrices provided they do not possess a zero eigenvalue.

From the above considerations, it is evident that the contributions $d_\Lambda P_S(t)/dt$ and $d_{\dot{O}} P_S(t)/dt$ to the rate of entropy production $dP_S(t)/dt$ for the case of discrete composite systems can be written as

$$\frac{d_\Lambda P_S(t)}{dt} = \sum_{\alpha=1}^{2} \frac{d_{\Lambda^{(\alpha)}} P_S^{(\alpha)}(t)}{dt} \tag{370}$$

and

$$\frac{d_{\dot{O}} P_S(t)}{dt} = \sum_{\alpha=1}^{2} \frac{d_{\dot{O}^{(\alpha)}} P_S^{(\alpha)}(t)}{dt}, \tag{371}$$

where

$$\frac{d_{\Lambda^{(\alpha)}} P_S^{(\alpha)}(t)}{dt} = k_B \sum_j \left[\frac{d\Lambda_j^{(\alpha)}(t)}{dt}\right]\left[\frac{d\langle \hat{O}_j^{(\alpha)}(t)\rangle}{dt}\right] \tag{372}$$

and

$$\frac{d_{\dot{O}^{(\alpha)}} P_S^{(\alpha)}(t)}{dt} = k_B \sum_j \Lambda_j^{(\alpha)}(t)\left[\frac{d^2\langle \hat{O}_j^{(\alpha)}(t)\rangle}{dt^2}\right], \tag{373}$$

with

$$\frac{d\langle \hat{O}_j^{(\alpha)}(t)\rangle}{dt} = -\sum_k \chi_{jk}^{(\alpha\alpha)}(t)\left[\frac{d\Lambda_k^{(\alpha)}(t)}{dt}\right] \tag{374}$$

and

$$\frac{d\Lambda_j^{(\alpha)}(t)}{dt} = -\sum_k \chi_{jk}^{(\alpha\alpha)^{-1}}(t)\left[\frac{d\langle \hat{O}_k^{(\alpha)}(t)\rangle}{dt}\right]. \tag{375}$$

In view of Eqs. (372), (374), and (375), we can write

$$\frac{d_{\Lambda^{(\alpha)}} P_S^{(\alpha)}(t)}{dt} = -k_B \sum_{j,k}\left[\frac{d\langle \hat{O}_j^{(\alpha)}(t)\rangle}{dt}\right]\chi_{jk}^{(\alpha\alpha)^{-1}}(t)\left[\frac{d\langle \hat{O}_k^{(\alpha)}(t)\rangle}{dt}\right] \tag{376a}$$

$$= -k_B \sum_{j,k}\left[\frac{d\Lambda_j^{(\alpha)}(t)}{dt}\right]\chi_{jk}^{(\alpha\alpha)}(t)\left[\frac{d\Lambda_k^{(\alpha)}(t)}{dt}\right]. \tag{376b}$$

Since $\boldsymbol{\chi}^{(\alpha\alpha)}(t)$ and $\boldsymbol{\chi}^{(\alpha\alpha)^{-1}}(t)$ are positive definite, these relations imply

$$\frac{d_{\Lambda^{(\alpha)}} P_S^{(\alpha)}(t)}{dt} \le 0. \tag{377}$$

This result reveals that the general evolution theorem given by Eq. (367) applies to both the composite system and its subsystems.

Let us turn our attention to fluctuations about a nonequilibrium state for our general quantum system. For sufficiently small fluctuations $\delta\langle \hat{O}_j(t)\rangle = \langle \hat{O}_j(t)\rangle_F - \langle \hat{O}_j(t)\rangle$ about the nonequilibrium state $\hat{\rho}(t)$ with the entropy $S(t)$, we can write the entropy $S_F(t)$ of the macroscopic state with the values $\{\langle \hat{O}_j(t)\rangle_F\}$ for the thermodynamic coordinates as

$$S_F(t) = S(t) + \delta S(t) + \frac{1}{2}\delta^2 S(t) + \cdots, \tag{378}$$

where the first and second variations in the entropy are given by

$$\delta S(t) = \sum_j \left[\frac{\partial S(t)}{\partial \langle \hat{O}_j(t)\rangle}\right] \delta\langle \hat{O}_j(t)\rangle \tag{379}$$

and

$$\delta^2 S(t) = \sum_{j,k} \left[\frac{\partial^2 S(t)}{\partial \langle \hat{O}_j(t)\rangle\, \partial \langle \hat{O}_k(t)\rangle}\delta\right] \delta\langle \hat{O}_j(t)\rangle \delta\langle \hat{O}_k(t)\rangle. \tag{380}$$

Making use of Eqs. (345) and (359), we rewrite Eqs. (379) and (380) as

$$\delta S(t) = k_{\rm B} \sum_j \Lambda_j(t) \delta\langle \hat{O}_j(t)\rangle \tag{381}$$

and

$$\delta^2 S(t) = k_{\rm B} \sum_{j,k} \delta\langle \hat{O}_j(t)\rangle \left[\frac{\partial \Lambda_j(t)}{\partial \langle \hat{O}_k(t)\rangle}\right] \delta\langle \hat{O}_k(t)\rangle \tag{382a}$$

$$= -k_{\rm B} \sum_{j,k} \delta\langle \hat{O}_j(t)\rangle\, \chi^{-1}_{jk}(t) \delta\langle \hat{O}_k(t)\rangle. \tag{382b}$$

Since $\boldsymbol{\chi}^{-1}(t)$ is positive definite, we conclude from Eq. (382b) that the inequality

$$\delta^2 S(t) < 0 \tag{383}$$

holds for nonvanishing fluctuations. Then

$$k_{\rm B} \sum_{j,k} \delta\langle \hat{O}_j(t)\rangle \left[\frac{\partial \Lambda_j(t)}{\partial \langle \hat{O}_k(t)\rangle}\right] \delta\langle \hat{O}_k(t)\rangle < 0 \tag{384}$$

or

$$-k_{\rm B} \sum_{j,k} \delta\langle \hat{O}_j(t)\rangle\, \chi^{-1}_{jk}(t)\, \delta\langle \hat{O}_k(t)\rangle < 0. \tag{385}$$

For the case of a closed homogeneous nonequilibrium system undergoing chemical reaction at a uniform temperature, the inequality (384) assumes the form of the stability condition

$$\sum_{j,k} \delta\xi_j(t) \left[\frac{\partial A_j(t)}{\partial \xi_k(t)}\right] \delta\xi_k(t) < 0 \tag{386}$$

postulated by Prigogine in nonequilibrium thermodynamics for chemically reactive systems. In the above, $\xi_j(t)$ denotes the progress variable for reaction j with the chemical affinity $A_j(t)$.

For discrete composite systems, we can write

$$S(t) = \sum_{\alpha=1}^{2} S^{(\alpha)}(t), \tag{387}$$

$$\delta S(t) = \sum_{\alpha=1}^{2} \delta S^{(\alpha)}(t), \tag{388}$$

and

$$\delta^2 S(t) = \sum_{\alpha=1}^{2} \delta^2 S^{(\alpha)}(t), \tag{389}$$

where

$$S^{(\alpha)}(t) = -k_{\rm B}\Omega^{(\alpha)}(t) + k_{\rm B} \sum_j \Lambda_j^{(\alpha)}(t) \langle \hat{O}_j^{(\alpha)}(t)\rangle, \tag{390}$$

$$\delta S^{(\alpha)}(t) = k_{\rm B} \sum_j \Lambda_j^{(\alpha)} \delta\langle \hat{O}_j^{(\alpha)}(t)\rangle, \tag{391}$$

and

$$\delta^2 S^{(\alpha)}(t) = k_{\rm B} \sum_{j,k} \delta\langle \hat{O}_j^{(\alpha)}(t)\rangle \left[\frac{\partial \Lambda_j^{(\alpha)}(t)}{\partial \langle \hat{O}_k^{(\alpha)}(t)\rangle}\right] \delta\langle \hat{O}_k^{(\alpha)}(t)\rangle \tag{392a}$$

$$= -k_{\rm B} \sum_{j,k} \delta\langle \hat{O}_j^{(\alpha)}(t)\rangle\, \chi^{(\alpha\alpha)^{-1}}_{jk}(t)\, \delta\langle \hat{O}_k^{(\alpha)}(t)\rangle. \tag{392b}$$

Thus,

$$S_F(t) = \sum_{\alpha=1}^{2} S_F^{(\alpha)}(t), \tag{393}$$

where

$$S_F^{(\alpha)}(t) = S^{(\alpha)}(t) + \delta S^{(\alpha)}(t) + \frac{1}{2}\delta^2 S^{(\alpha)}(t) + \cdots. \tag{394}$$

Since $\boldsymbol{\chi}^{(\alpha\alpha)^{-1}}(t)$ is positive definite, we conclude from Eq. (392b) that the inequality

$$\delta^2 S^{(\alpha)}(t) < 0 \tag{395}$$

holds for nonvanishing fluctuations. Then

$$k_{\rm B} \sum_{j,k} \delta\langle \hat{O}_j^{(\alpha)}(t)\rangle \left[\frac{\partial \Lambda_j^{(\alpha)}(t)}{\partial \langle \hat{O}_k^{(\alpha)}(t)\rangle}\right] \delta\langle \hat{O}_k^{(\alpha)}(t)\rangle < 0 \tag{396}$$

or

$$-k_{\rm B} \sum_{j,k} \delta\langle \hat{O}_j^{(\alpha)}(t)\rangle\, \chi^{(\alpha\alpha)^{-1}}_{jk}(t)\, \delta\langle \hat{O}_k^{(\alpha)}(t)\rangle < 0. \tag{397}$$

Each subsystem of the discrete composite system must satisfy (395)–(397). The discrete composite system itself must satisfy (383)–(385).

If the quantities appearing in (384) and (385) are time independent, we obtain

$$k_{\mathrm{B}} \sum_{j,k} \delta\langle \hat{O}_j \rangle \left[\frac{\partial \Lambda_j}{\partial \langle \hat{O}_k \rangle} \right] \delta\langle \hat{O}_k \rangle < 0 \tag{398}$$

or

$$-k_{\mathrm{B}} \sum_{j,k} \delta\langle \hat{O}_j \rangle \chi_{jk}^{-1} \delta\langle \hat{O}_k \rangle < 0. \tag{399}$$

For the case of discrete composite systems, we can also write

$$k_{\mathrm{B}} \sum_{j,k} \delta\langle \hat{O}_j^{(\alpha)} \rangle \left[\frac{\partial \Lambda_j^{(\alpha)}}{\partial \langle \hat{O}_k^{(\alpha)} \rangle} \right] \delta\langle \hat{O}_k^{(\alpha)} \rangle < 0 \tag{400}$$

or

$$-k_{\mathrm{B}} \sum_{j,k} \delta\langle \hat{O}_j^{(\alpha)} \rangle \chi_{jk}^{(\alpha\alpha)^{-1}} \delta\langle \hat{O}_k^{(\alpha)} \rangle < 0 \tag{401}$$

for each subsystem α.

The inequality (400) is the Gibbs' stability condition for a subsystem in equilibrium thermodynamics. The result given by (398) may be regarded as a stability condition for the composite system.

In the above formal development, we found that Gibbs' stability condition from equilibrium thermodynamics and Prigogine's stability condition from nonequilibrium thermodynamics for a chemically reactive system emerge from the statistical mechanical treatment of nonequilibrium systems. Unlike the stability conditions of Gibbs and Prigogine, the inequalities (384), (385), (396), and (397) are not postulates. They are simple consequences of the statistical mechanical treatment. Moreover, these inequalities apply to both equilibrium and nonequilibrium systems.

The validity of the inequalities (384) and (385) does not require a physical system to be decomposable into two uncorrelated systems as in the case of discrete composite systems. This flexibility affords us with the opportunity to deal with physical situations not traditionally considered in equilibrium and nonequilibrium thermodynamics.

In view of the relation (359), the partial derivatives $\partial \Lambda_j(t)/\partial \langle \hat{O}_k(t) \rangle$ can become very small when the fluctuations, as described by the matrix elements $\chi_{jk}(t)$ of $\boldsymbol{\chi}(t)$, become very large. [This can also lead to very small negative values for $\delta^2 S(t)$.] In equilibrium thermodynamics, such behavior is said to signal the onset of a phase transition or the formation of a critical state of matter. Our statistical mechanical treatment reveals that this kind of behavior can be realized in both equilibrium and nonequilibrium systems.

An equilibrium phase transition is usually viewed as an abrupt change from an equilibrium state with a well-defined spatial order to another equilibrium state with a differing well-defined spatial order. In contrast, nonequilibrium phase transitions are manifested as an abrupt change from a nonequilibrium state with a well-defined spatiotemporal order to another nonequilibrium state with a differing well-defined spatiotemporal order.

An example of a nonequilibrium phase transition is the abrupt passage between the low- and high-transmission states in optically bistable systems. For phase transitions of this type, both the initial and final states are nonstationary (time-dependent) nonequilibrium states. In fact, the initial and final states display temporal oscillations.

Some nonequilibrium phase transitions involve the passage of a system from a stationary (time-independent) nonequilibrium state to a nonstationary nonequilibrium state with rich spatiotemporal order. For example, it has been observed that certain chemically reactive systems can pass from a quiescent homogeneous state to a state characterized by spatial and/or temporal oscillations in the concentrations of certain chemical species. In nonequilibrium thermodynamics, Prigogine has argued that such phase transitions result from the violation of the stability condition

$$\frac{1}{2}\frac{d}{dt}\delta^2 S(t) = \sum_k \delta J_k(t) \delta X_k(t) > 0. \tag{402}$$

In general, the stability condition given by (402) applies only for small fluctuations about a stationary state. Thus, its applicability is limited to a rather restricted class of physical phenomena, not including the aforementioned phenomenon of optical bistability.

One can demonstrate that the stability condition given by (402) follows from Eqs. (389) and (392b) provided (i) the matrix elements $\chi_{jk}^{(\alpha\alpha)}(t)$ of the susceptibility submatrices $\boldsymbol{\chi}^{(\alpha\alpha)}(t)$ are time independent, and (ii) the requirements of a Lyapounov function are imposed upon $-\delta^2 S(t)$. Since $\hat{\bar{\rho}}(t)$ is time independent for a stationary state, the matrix elements $\chi_{jk}^{(\alpha\alpha)}(t)$ of the susceptibility submatrices $\boldsymbol{\chi}^{(\alpha\alpha)}(t)$ are time independent. With the exception of systems exhibiting small fluctuations about a stationary state, $-\delta^2 S(t)$ cannot, in general, be regarded as a Lyapounov function.

The first variation $\delta S(t)$ in the entropy, given by Eq. (381), vanishes only when the fluctuations $\delta\langle \hat{O}_j(t) \rangle$ are about the state of maximum realizable entropy for all conceivable values $\{\langle \hat{O}_j(t) \rangle_c\}$ of the thermodynamic coordinates $\{\langle \hat{O}_j(t) \rangle\}$, i.e., the state of equilibrium. This may or may not be the state described by $\hat{\bar{\rho}}(t)$. In general, $\hat{\bar{\rho}}(t)$ is the state of maximum entropy only for a given set of values $\{\langle \hat{O}_j(t) \rangle_v\}$ of the thermodynamic coordinates $\{\langle \hat{O}_j(t) \rangle\}$. For nonequilibrium systems, the first variation $\delta S(t)$ in the entropy does not vanish.

The aforementioned conditions for a system in equilibrium or nonequilibrium can be expressed as follows:

$$\begin{aligned} \delta S(t) &= 0 \qquad \text{equilibrium} \\ \delta S(t) &\neq 0 \qquad \text{nonequilibrium} \end{aligned} \tag{403}$$

or

$$\begin{aligned} k_{\mathrm{B}} \sum_j \Lambda_j(t)\, \delta\langle \hat{O}_j(t)\rangle &= 0 \qquad \text{equilibrium} \\ k_{\mathrm{B}} \sum_j \Lambda_j(t)\, \delta\langle \hat{O}_j(t)\rangle &\neq 0 \qquad \text{nonequilibrium} \end{aligned} \tag{404}$$

The conditions (404) assume the following forms for discrete composite systems:

$$\begin{aligned} &k_{\mathrm{B}} \sum_j \left[\Lambda_j^{(1)}(t) - \Lambda_j^{(2)}(t)\right] \delta\left\langle \hat{O}_j^{(1)}(t)\right\rangle = 0 \quad \text{equilibrium} \\ &k_{\mathrm{B}} \sum_j \left[\Lambda_j^{(1)}(t) - \Lambda_j^{(2)}(t)\right] \delta\left\langle \hat{O}_j^{(1)}(t)\right\rangle \neq 0 \quad \text{nonequilibrium} \end{aligned} \tag{405}$$

As in the statistical mechanical and thermodynamic treatments of equilibrium systems, we find that the equilibrium condition implies $\Lambda_j^{(1)}(t) = \Lambda_j^{(2)}(t)$.

Since the relations $\delta S(t) = 0$ and $\delta^2 S(t) < 0$ hold for equilibrium systems, fluctuations $\delta\langle \hat{O}_j(t)\rangle$ about an equilibrium state $\hat{\bar{\rho}}(t) = \hat{\rho}_{\mathrm{eq}}$ bring about a decrease in the entropy, i.e., $S_F(t) < S(t)$ for sufficiently small fluctuations. For the case of nonequilibrium systems, the relations $\delta S(t) \neq 0$ and $\delta^2 S(t) < 0$ hold. Thus, fluctuations $\delta\langle \hat{O}_j(t)\rangle$ about a nonequilibrium state $\hat{\bar{\rho}}(t)$ can bring about an increase or a decrease in the entropy, i.e., $S_F(t) < S(t)$ or $S_F(t) > S(t)$.

Neglecting third-order and higher order contributions on the right side of Eq. (378), we write

$$\begin{aligned} S_F(t) - S(t) = k_{\mathrm{B}} \sum_j \Lambda_j(t) \delta\langle \hat{O}_j(t)\rangle - k_{\mathrm{B}} \sum_{j,k} \delta \\ \times \langle \hat{O}_j(t) \chi_{jk}^{-1}(t) \delta\langle \hat{O}_k(t)\rangle \end{aligned} \tag{406}$$

or

$$\begin{aligned} S_F(t) - S(t) = k_{\mathrm{B}} \sum_j \Lambda_j(t) \delta\langle \hat{O}_j(t)\rangle \\ + k_{\mathrm{B}} \sum_{j,k} \delta\langle \hat{O}_j(t)\rangle \left[\frac{\partial \Lambda_j(t)}{\partial \langle \hat{O}_k(t)\rangle}\right] \delta\langle \hat{O}_k(t)\rangle. \end{aligned} \tag{407}$$

For discrete composite systems, these equations can be written

$$S_F(t) - S(t) = \sum_{\alpha=1}^{2} \left[S_F^{(\alpha)}(t) - S^{(\alpha)}(t)\right], \tag{408}$$

where

$$\begin{aligned} S_F^{(\alpha)}(t) - S^{(\alpha)}(t) = k_{\mathrm{B}} \sum_j \Lambda_j^{(\alpha)} \delta\left\langle \hat{O}_j^{(\alpha)}(t)\right\rangle \\ - k_{\mathrm{B}} \sum_{j,k} \delta\left\langle \hat{O}_j^{(\alpha)}(t)\right\rangle \chi_{jk}^{(\alpha\alpha)^{-1}}(t) \delta\left\langle \hat{O}_k^{(\alpha)}(t)\right\rangle \end{aligned} \tag{409}$$

or

$$\begin{aligned} S_F^{(\alpha)}(t) - S^{(\alpha)}(t) = k_{\mathrm{B}} \sum_j \Lambda_j^{(\alpha)} \delta\left\langle \hat{O}_j^{(\alpha)}(t)\right\rangle \\ + k_{\mathrm{B}} \sum_{j,k} \delta\left\langle \hat{O}_j^{(\alpha)}(t)\right\rangle \left[\frac{\partial \Lambda_j^{(\alpha)}(t)}{\partial \left\langle \hat{O}_k^{(\alpha)}(t)\right\rangle}\right] \delta\left\langle \hat{O}_k^{(\alpha)}(t)\right\rangle. \end{aligned} \tag{410}$$

If the fluctuations $\delta\langle \hat{O}_j(t)\rangle$ are about an equilibrium state $\hat{\bar{\rho}}(t) = \hat{\rho}_{\mathrm{eq}}$ with the entropy $S(t) = S_{\mathrm{eq}}$, we find that Eq. (406) assumes the form

$$S_F(t) - S_{\mathrm{eq}} = -k_{\mathrm{B}} \sum_{j,k} \delta\langle \hat{O}_j(t)\rangle\, \chi_{jk}^{-1} \delta\langle \hat{O}_k(t)\rangle, \tag{411}$$

where χ_{jk}^{-1} is a matrix element of the inverse $\boldsymbol{\chi}^{-1}$ of the equilibrium susceptibility matrix $\boldsymbol{\chi}$. The equilibrium susceptibilities χ_{jk} are given by

$$\chi_{jk} = \langle \delta\hat{O}_k \delta\hat{\bar{O}}_j\rangle_{\mathrm{eq}} \tag{412a}$$

$$= \mathrm{Tr}\, \hat{\rho}_{\mathrm{eq}} \delta\hat{O}_k \delta\hat{\bar{O}}_j, \tag{412b}$$

where

$$\delta\hat{\bar{O}}_j = \hat{\bar{O}}_j - \langle \hat{\bar{O}}_j\rangle_{\mathrm{eq}} \tag{413}$$

and

$$\delta\hat{O}_k = \hat{O}_k - \langle \hat{O}_k\rangle_{\mathrm{eq}}, \tag{414}$$

with

$$\delta\hat{\bar{O}}_j = \int_0^1 d\lambda\, \hat{\rho}_{\mathrm{eq}}^{\lambda} \hat{O}_j \hat{\rho}_{\mathrm{eq}}^{-\lambda} \tag{415}$$

and

$$\langle \hat{O}_j\rangle_{\mathrm{eq}} = \mathrm{Tr}\, \hat{\rho}_{\mathrm{eq}} \hat{O}_j. \tag{416}$$

The result given by Eq. (411) for $S_F(t)$ is the expression assumed for the entropy $S(t)$ in linear nonequilibrium thermodynamics on the basis of Einstein's theory of fluctuations. In nonequilibrium thermodynamics, it is used to justify the expressions for the entropy production and thermodynamic driving forces. On the basis of the statistical mechanical treatment of nonequilibrium systems, we find that this approximate approach is not required.

In a modern approach to thermodynamic fluctuations, Callen expresses the instantaneous entropy S_I for the thermodynamic state arising from fluctuations δO_j about the thermodynamic state with the entropy S as

$$S_I - S = \delta S + \delta^2 S + \cdots, \tag{417}$$

where

$$\delta S = \sum_j X_j \delta O_j \tag{418}$$

and

$$\delta^2 S = -\frac{1}{2}\sum_{j,k} \delta O_j g_{jk}\, \delta O_k, \tag{419}$$

with

$$g_{jk} = -\frac{\partial^2 S}{\partial O_j \partial O_k}. \tag{420}$$

More explicitly,

$$S_I - S = \sum_j X_j \delta O_j - \frac{1}{2}\sum_{j,k} \delta O_j g_{jk} \delta O_k + \cdots \tag{421}$$

In Callen's formulation of thermodynamic fluctuations, it is found that

$$k_{\rm B} g_{jk}^{-1} = \langle \delta O_j \delta O_k \rangle_f \tag{422a}$$

$$= \int d\delta O_1 \cdots \int d\delta O_n\, f(\delta O_1, \ldots, \delta O_n) \delta O_j \delta O_k, \tag{422b}$$

where

$$f(\delta O_1, \ldots, \delta O_n) = f_0 \exp\left(+\frac{1}{2}\delta^2 S + \cdots\right) \tag{423}$$

represents a probability density for the fluctuations. It is also shown that

$$k_{\rm B} g_{jk}^{-1} = -\frac{\partial O_j}{\partial X_k} = -\frac{\partial O_k}{\partial X_j}. \tag{424}$$

Provided third-order and higher variations in the entropy are neglected, the results given by Eqs. (422a)–(424) are equivalent to the results given by Einstein's theory of fluctuations. Nonetheless, Callen's formulation is more general than the theory given by Einstein.

In Einstein's theory of fluctuations, the linear term in the expression for the instantaneous entropy S_I, given by Eq. (421), is absent due to the requirement that the fluctuations be executed about an equilibrium state. In the Callen treatment of thermodynamic fluctuations, the entropy S is the entropy of the most probable state, which does not necessarily represent an equilibrium state. Hence, the linear term in Eq. (421) is, in general, nonvanishing.

The entropies S_I and S in Callen's formulation of thermodynamic fluctuations bear a striking resemblance to the entropies $S_F(t)$ and $S(t)$ in the statistical mechanical treatment of fluctuations about a nonequilibrium state. Making the identifications $X_j(t) = k_{\rm B}\Lambda_j(t)$ and $\delta O_j = \delta\langle \hat{O}_j(t)\rangle$, we conclude that $S_I = S_F(t)$ and $S = S(t)$. Also, we find that the matrix elements g_{jk}^{-1} of the inverse $\mathbf{g}^{-1}$ of the matrix $\mathbf{g}$ are given by $k_{\rm B} g_{jk}^{-1} = \chi_{jk}(t)$ provided $\langle \delta O_k \delta O_j \rangle_f = \langle \delta \hat{O}_k(t) \delta \hat{\bar{O}}_j(t)\rangle_{\bar{\rho}(t)}$. In addition, we find that Eq. (424) is equivalent to the mathematical statements given by Eqs. (350) and (358).

Unlike the theories of thermodynamic fluctuations formulated by Einstein and Callen, the statistical mechanical treatment provides a connection between the fluctuations and microscopic interactions through the formal expression given by Eqs. (360a)–(363) for the susceptibilities $\chi_{jk}(t)$.

2. Spatially Dependent Thermodynamic Coordinates

Making use of the time-dependent version of the maximum entropy principle, we find that the statistical density operator $\hat{\bar{\rho}}(t)$ for a system characterized by the spatially dependent thermodynamic coordinates $\{\langle \hat{O}_j(\vec{r}, t)\rangle\}$ is given by

$$\hat{\bar{\rho}}(t) = \exp\left[\Omega(t)\hat{I} - \sum_j \int_{\mathcal{D}} dV \Lambda_j(\vec{r}, t)\hat{O}_j(\vec{r})\right], \tag{425}$$

where

$$\Omega(t) = -\ln Z(t), \tag{426}$$

with

$$Z(t) = {\rm Tr} \exp\left[-\sum_j \int_{\mathcal{D}} dV \Lambda_j(\vec{r}, t)\hat{O}_j(\vec{r})\right]. \tag{427}$$

In view of Eq. (340), we can write the entropy $S(t)$ of the macroscopic state described by $\hat{\bar{\rho}}(t)$ as

$$S(t) = \Phi(t) + k_{\rm B}\sum_j \int_{\mathcal{D}} dV \Lambda_j(\vec{r}, t)\langle \hat{O}_j(\vec{r}, t)\rangle, \tag{428}$$

where $\Phi(t) = -k_{\rm B}\Omega(t)$. The entropy $S(t)$ is a functional $S[O_t]$ of the thermodynamic coordinates $O_t = \{\langle \hat{O}_j(\vec{r}, t)\rangle\}$. Similarly, the thermodynamic potential $\Phi(t)$ is a functional $\Phi[\Lambda_t]$ of the thermodynamic parameters $\Lambda_t = \{\Lambda_j(\vec{r}, t)\}$.

One can demonstrate that

$$\delta S(t) = k_{\rm B}\sum_j \int_{\mathcal{D}} dV \Lambda_j(\vec{r}, t)\delta\langle \hat{O}_j(\vec{r}, t)\rangle \tag{429}$$

and

$$\delta\Phi(t) = -k_{\rm B}\sum_j \int_{\mathcal{D}} dV \langle \hat{O}_j(\vec{r}, t)\rangle \delta\Lambda_j(\vec{r}, t). \tag{430}$$

In the above, $\delta S(t) = S[O_t + \delta O_t] - S[O_t]$ is the variation in the entropy $S(t)$ brought about by the variations

$\delta O_t = \{\delta\langle\hat{O}_j(\vec{r},t)\rangle\}$ in the thermodynamic coordinates. Similarly, $\delta\Phi(t) = \Phi[\Lambda_t + \delta\Lambda_t] - \Phi[\Lambda_t]$ is the variation in the thermodynamic potential $\Phi(t)$ brought about by the variations $\delta\Lambda_t = \{\delta\Lambda_j(\vec{r},t)\}$ in the thermodynamic parameters.

It follows from Eqs. (429) and (430) that the thermodynamic parameters and thermodynamic coordinates are given by the functional derivatives

$$\frac{\delta S(t)}{\delta\langle\hat{O}_j(\vec{r},t)\rangle} = k_B\Lambda_j(\vec{r},t) \tag{431}$$

and

$$\frac{\delta\Phi(t)}{\delta\Lambda_j(\vec{r},t)} = -k_B\langle\hat{O}_j(\vec{r},t)\rangle. \tag{432}$$

The time dependence of the entropy $S(t)$ is due to its dependence on the thermodynamic coordinates. Making use of this observation and Eq. (431), we find that the entropy production $P_S(t)$ is given by

$$P_S(t) = k_B\sum_j\int_{\mathcal{D}} dV\,\Lambda_j(\vec{r},t)\,\frac{\partial\langle\hat{O}_j(\vec{r},t)\rangle}{\partial t}. \tag{433}$$

For the case in which the thermodynamic coordinates $\{\langle\hat{O}_j(\vec{r},t)\rangle\}$ represent the local densities $\{\rho_j(\vec{r},t)\}$ of conserved quantities, the above equation is equivalent to the bilinear form implicitly assumed for the entropy production in nonequilibrium thermodynamics for continuous systems. [See Eq. (303).]

Assuming that the thermodynamic coordinates $\{\langle\hat{O}_j(\vec{r},t)\rangle\}$ do indeed represent the local densities $\{\rho_j(\vec{r},t)\}$ of conserved quantities, we can write Eqs. (428)–(430) as

$$S(t) = \Phi(t) + k_B\sum_j\int_{\mathcal{D}} dV\,\Lambda_j(\vec{r},t)\rho_j(\vec{r},t), \tag{434}$$

$$\delta S(t) = k_B\sum_j\int_{\mathcal{D}} dV\,\Lambda_j(\vec{r},t)\delta\rho_j(\vec{r},t), \tag{435}$$

and

$$\delta\Phi(t) = -k_B\sum_j\int_{\mathcal{D}} dV\rho_j(\vec{r},t)\delta\Lambda_j(\vec{r},t). \tag{436}$$

The entropy density $\rho_s(\vec{r},t)$ and the thermodynamic potential density $\rho_\Phi(\vec{r},t)$ can be introduced through the relations

$$S(t) = \int_{\mathcal{D}} dV\rho_s(\vec{r},t) \tag{437}$$

and

$$\Phi(t) = \int_{\mathcal{D}} dV\rho_\Phi(\vec{r},t). \tag{438}$$

Making use of these definitions and Eqs. (434)–(436), we write

$$\rho_s(\vec{r},t) = \rho_\Phi(\vec{r},t) + k_B\sum_j\Lambda_j(\vec{r},t)\rho_j(\vec{r},t), \tag{439}$$

$$\delta\rho_s(\vec{r},t) = k_B\sum_j\Lambda_j(\vec{r},t)\delta\rho_j(\vec{r},t), \tag{440}$$

and

$$\delta\rho_\Phi(\vec{r},t) = -k_B\sum_j\rho_j(\vec{r},t)\,\delta\Lambda_j(\vec{r},t). \tag{441}$$

The above expression for the variation $\delta\rho_s(\vec{r},t)$ in the entropy density $\rho_s(\vec{r},t)$ is the variational analogue of the Gibbsian expression assumed for the differential $d\rho_s(\vec{r},t)$ of the entropy density $\rho_s(\vec{r},t)$ in nonequilibrium thermodynamics. [See Eq. (297).]

As indicated earlier, the expression for the entropy production $P_S(t)$ in nonequilibrium thermodynamics, given by Eq. (303) follows from Eq. (433). Making use of this expression, the continuity equations $\partial\rho_j(\vec{r},t)/\partial t = -\vec{\nabla}\cdot\vec{J}_j(\vec{r},t)$, and Gauss' theorem, we obtain Eq. (292) for the entropy production in nonequilibrium thermodynamics with the contributions $P_S^{(i)}(t)$ and $P_S^{(e)}(t)$ given by Eqs. (304) and (305), respectively. As we have demonstrated, these basic results from nonequilibrium thermodynamics follow directly from the statistical mechanical treatment of nonequilibrium systems without introducing the notion of a local entropy density $\rho_s(\vec{r},t)$ whose time evolution is governed by an entropy balance equation.

Making use of Eq. (433), we write the rate of entropy production as

$$\frac{dP_S(t)}{dt} = \frac{d_\Lambda P_S(t)}{dt} + \frac{d_{\dot{O}}P_S(t)}{dt}, \tag{442}$$

where

$$\frac{d_\Lambda P_S(t)}{dt} = k_B\sum_j\int_{\mathcal{D}} dV\left[\frac{\partial\Lambda_j(\vec{r},t)}{\partial t}\right]\left[\frac{\partial\langle\hat{O}_j(\vec{r},t)\rangle}{\partial t}\right] \tag{443}$$

and

$$\frac{d_{\dot{O}}P_S(t)}{dt} = k_B\sum_j\int_{\mathcal{D}} dV\,\Lambda_j(\vec{r},t)\left[\frac{\partial^2\langle\hat{O}_j(\vec{r},t)\rangle}{\partial t^2}\right]. \tag{444}$$

If the thermodynamic coordinates $\{\langle\hat{O}_j(\vec{r},t)\rangle\}$ represent local densities $\{\rho_j(\vec{r},t)\}$ of conserved quantities, one can demonstrate that these equations can be recast in the forms given by Eqs. (312)–(318) from nonequilibrium thermodynamics.

One can show that

$$\frac{\delta\langle\hat{O}_j(\vec{r},t)\rangle}{\delta\Lambda_k(\vec{r}',t)} = -\chi_{jk}(\vec{r},\vec{r}';t) \tag{445}$$

and

$$\frac{\delta\Lambda_j(\vec{r},t)}{\delta\langle\hat{O}_k(\vec{r}',t)\rangle} = -\chi^{-1}_{jk}(\vec{r},\vec{r}';t). \tag{446}$$

In the above, $\chi^{-1}_{jk}(\vec{r},\vec{r}';t)$ denotes a matrix element of the inverse $\chi^{-1}(t)$ of the matrix $\chi(t)$. The inverse $\chi^{-1}(t)$ of $\chi(t)$ is defined in such a way that

$$\sum_l \int_{\mathcal{D}} dV''\chi_{jl}(\vec{r},\vec{r}'')\,\chi_{lk}(\vec{r}'',\vec{r}') = \delta_{jk}\,\delta(\vec{r}-\vec{r}').$$

The matrix elements $\chi_{jk}(\vec{r},\vec{r}';t)$ of $\chi(t)$ are the generalized nonequilibrium susceptibilities

$$\chi_{jk}(\vec{r},\vec{r}';t) = \langle\delta_t\hat{O}_k(\vec{r}')\delta_t\hat{\hat{O}}_j(\vec{r}\,)\rangle_{\hat{\rho}(t)} \tag{447a}$$

$$= \mathrm{Tr}\,\hat{\rho}(t)\,\delta_t\hat{O}_k(\vec{r}')\,\delta_t\hat{\hat{O}}_j(\vec{r}\,), \tag{447b}$$

where

$$\delta_t\hat{\hat{O}}_j(\vec{r}) = \hat{\hat{O}}_j(\vec{r}) - \langle\hat{\hat{O}}_j(\vec{r},t)\rangle \tag{448}$$

and

$$\delta_t\hat{O}_k(\vec{r}') = \hat{O}_k(\vec{r}') - \langle\hat{O}_k(\vec{r}',t)\rangle, \tag{449}$$

with

$$\hat{\hat{O}}_j(\vec{r}) = \int_0^1 d\lambda\,\hat{\rho}(t)^\lambda\hat{O}_j(\vec{r})\,\hat{\rho}(t)^{-\lambda}. \tag{450}$$

In general, $\langle\hat{\hat{O}}_j(\vec{r},t)\rangle = \langle\hat{O}_j(\vec{r},t)\rangle$.

Making use of Eqs. (447a)–(450), one can demonstrate that the matrix $\chi(t)$ possesses the following properties:

1. $\chi_{jk}(\vec{r},\vec{r}';t) = \chi_{kj}(\vec{r}',\vec{r};t)$.
2. $\chi^*_{jk}(\vec{r},\vec{r}';t) = \chi_{jk}(\vec{r},\vec{r}';t)$.
3. $\sum_{j,k}\int_{\mathcal{D}} dV\int_D dV'\Phi^*_j(\vec{r},t)\chi_{jk}(\vec{r},\vec{r}';t)\Phi_k(\vec{r}',t) \geq 0$ for any set of complex functions $\{\Phi_j(\vec{r},t)\}$, and $\sum_{j,k}\int_{\mathcal{D}} dV\int_{\mathcal{D}} dV'\phi_j(\vec{r},t)\chi_{jk}(\vec{r},\vec{r}';t)\phi_k(\vec{r}',t) \geq 0$ for any set of real functions $\{\phi_j(\vec{r},t)\}$.

Thus, $\chi(t)$ is a real symmetric, nonnegative-definite matrix. Assuming that $\chi(t)$ does not possess a zero eigenvalue, $\chi(t)$ and $\chi^{-1}(t)$ are positive definite.

Treating $\langle\hat{O}_j(\vec{r},t)\rangle$ as a functional $O_j[\vec{r},\Lambda_t]$ of the thermodynamic parameters $\Lambda_t = \{\Lambda_j(\vec{r},t)\}$, and $\Lambda_j(\vec{r},t)$ as a functional $\Lambda_j[\vec{r},O_t]$ of the thermodynamic coordinates $O_t = \{\langle\hat{O}_j(\vec{r},t)\rangle\}$, one can make use of Eqs. (445) and (446) to show that

$$\frac{\partial\langle\hat{O}_j(\vec{r},t)\rangle}{\partial t} = -\sum_k\int_{\mathcal{D}} dV'\chi_{jk}(\vec{r},\vec{r}';t)\,\frac{\partial\Lambda_k(\vec{r}',t)}{\partial t} \tag{451}$$

and

$$\frac{\partial\Lambda_j(\vec{r},t)}{\partial t} = -\sum_k\int_{\mathcal{D}} dV'\chi^{-1}_{jk}(\vec{r},\vec{r}';t)\,\frac{\partial\langle\hat{O}_k(\vec{r}',t)\rangle}{\partial t}. \tag{452}$$

With these equations at our disposal, we can rewrite Eq. (443) as

$$\frac{d_\Lambda P_S(t)}{dt} = -k_{\mathrm{B}}\sum_{j,k}\int_{\mathcal{D}} dV\int_{\mathcal{D}} dV'\left[\frac{\partial\langle\hat{O}_j(\vec{r},t)\rangle}{\partial t}\right] \times\chi^{-1}_{jk}(\vec{r},\vec{r}';t)\left[\frac{\partial\langle\hat{O}_k(v\vec{r}',t)\rangle}{\partial t}\right] \tag{453a}$$

$$= -k_{\mathrm{B}}\sum_{j,k}\int_{\mathcal{D}} dV\int_{\mathcal{D}} dV'\left[\frac{\partial\Lambda_j(\vec{r},t)}{\partial t}\right] \times\chi_{jk}(\vec{r},\vec{r}';t)\left[\frac{\partial\Lambda_k(\vec{r}',t)}{\partial t}\right]. \tag{453b}$$

Since $\chi(t)$ and $\chi^{-1}(t)$ are positive definite, the relations (453a) and (453b) imply that the inequality (367) is satsified. As indicated earlier, this inequality is a generalization of the Prigogine–Glansdorff evolution theorem that applies independent of the nature of the physical system.

If the thermodynamic coordinates $\{\langle\hat{O}_j(\vec{r},t)\rangle\}$ represent the local densities $\{\rho_j(\vec{r},t)\}$ of conserved quantities, we can rewrite Eq. (367) as

$$\frac{d_X P^{(\mathrm{i})}_S(t)}{dt} + \frac{d_\Lambda P^{(\mathrm{e})}_S(t)}{dt} \leq 0, \tag{454}$$

where $d_X P^{(\mathrm{i})}_S(t)/dt$ and $d_\Lambda P^{(\mathrm{e})}_S(t)/dt$ are given by Eqs. (315) and (317), respectively. Assuming that thermodynamic parameters $\Lambda_j(\vec{r},t)$ on the boundary surface between the system and its surroundings are time independent, we recover the Glansdorff–Prigogine evolution theorem for continuous systems in nonequilibrium thermodynamics. [See Eq. (319).]

Unlike the Glansdorff–Prigogine treatment of nonequilibrium systems, the statistical mechanical treatment yields the evolution theorem given by Eq. (319) and the more general evolution theorem given by Eq. (367) without postulating the validity of Gibbs stability theory for spatially local systems. In fact, we demonstrate below that the fundamental relation required in the Glansdorff–Prigogine treatment emerges from the statistical mechanical treatment of nonequilibrium systems.

Let us turn our attention to fluctuations about a nonequilibrium state for our general quantum system. For sufficiently small fluctuatioins $\delta\langle\hat{O}_j(\vec{r},t)\rangle = \langle\hat{O}_j(\vec{r},t)\rangle_F - \langle\hat{O}_j(\vec{r},t)\rangle$ about the nonequilibrium state $\hat{\bar{\rho}}(t)$ with the entropy $S(t)$, we can write the entropy $S_F(t)$ of the macroscopic state with the values $\{\langle\hat{O}_j(\vec{r},t)\rangle_F\}$ for the thermodynamic coordinates $\{\langle\hat{O}_j(\vec{r},t)\rangle\}$ as

$$S_F(t) = S(t) + \delta S(t) + \frac{1}{2}\delta^2 S(t) + \cdots, \tag{455}$$

where the first and second variations in the entropy are given by

$$\delta S(t) = \sum_j \int_{\mathcal{D}} dV \left[\frac{\delta S(t)}{\delta\langle\hat{O}_j(\vec{r},t)\rangle}\right]\delta\langle\hat{O}_j(\vec{r},t)\rangle \tag{456}$$

and

$$\delta^2 S(t) = \sum_{j,k}\int_{\mathcal{D}} dV \int_{\mathcal{D}} dV' \left[\frac{\delta^2 S(t)}{\delta\langle\hat{O}_j(\vec{r},t)\rangle\,\delta\langle\hat{O}_k(\vec{r}',t)\rangle}\right] \times \delta\langle\hat{O}_j(\vec{r},t)\rangle\,\delta\langle\hat{O}_k(\vec{r}',t)\rangle. \tag{457}$$

Making use of Eqs. (431) and (446), we can rewrite Eqs. (456) and (457) as

$$\delta S(t) = k_{\mathrm{B}} \sum_j \int_{\mathcal{D}} dV\, \Lambda_j(\vec{r},t)\delta\langle\hat{O}_j(\vec{r},t)\rangle \tag{458}$$

and

$$\delta^2 S(t) = k_{\mathrm{B}} \sum_{j,k}\int_{\mathcal{D}} dV \int_{\mathcal{D}} dV'\, \delta\langle\hat{O}_j(\vec{r},t)\rangle \times \left[\frac{\delta\Lambda_j(\vec{r},t)}{\delta\langle\hat{O}_k(\vec{r}',t)\rangle}\right]\delta\langle\hat{O}_k(\vec{r}',t)\rangle \tag{459a}$$

$$= -k_{\mathrm{B}} \sum_{j,k}\int_{\mathcal{D}} dV \int_{\mathcal{D}} dV'\, \delta\langle\hat{O}_j(\vec{r},t)\rangle \times \chi^{-1}_{jk}(\vec{r},\vec{r}';t)\,\delta\langle\hat{O}_k(\vec{r}',t)\rangle. \tag{459b}$$

Since $\boldsymbol{\chi}^{-1}(t)$ is positive definite, we conclude from Eq. (459b) that the inequality

$$\delta^2 S(t) < 0 \tag{460}$$

holds for nonvanishing fluctuations. Then

$$k_{\mathrm{B}} \sum_{j,k}\int_{\mathcal{D}} dV \int_{\mathcal{D}} dV'\delta\langle\hat{O}_j(\vec{r},t)\rangle \times \left[\frac{\delta\Lambda_j(\vec{r},t)}{\delta\langle\hat{O}_k(\vec{r}',t)\rangle}\right]\delta\langle\hat{O}_k(\vec{r}',t)\rangle < 0 \tag{461}$$

or

$$-k_{\mathrm{B}} \sum_{j,k}\int_{\mathcal{D}} dV \int_{\mathcal{D}} dV'\, \delta\langle\hat{O}_j(\vec{r},t)\rangle \times \chi^{-1}_{jk}(\vec{r},\vec{r}';t)\,\delta\langle\hat{O}_k(\vec{r}',t)\rangle < 0. \tag{462}$$

If the thermodynamic coordinates $\{\langle\hat{O}_j(\vec{r},t)\rangle\}$ represent local densities $\{\rho_j(\vec{r},t)\}$ of conserved quantities, we can write

$$k_{\mathrm{B}} \sum_{j,k}\int_{\mathcal{D}} dV \int_{\mathcal{D}} dV'\delta\rho_j(\vec{r},t)\left[\frac{\delta\Lambda_j(\vec{r},t)}{\delta\rho_k(\vec{r}',t)}\right]\delta\rho_k(\vec{r}',t) < 0 \tag{463}$$

or

$$-k_{\mathrm{B}} \sum_{j,k}\int_{\mathcal{D}} dV \int_{\mathcal{D}} dV'\delta\rho_j(\vec{r},t)\chi^{-1}_{jk}(\vec{r},\vec{r}';t)\delta\rho_k(\vec{r}',t) < 0. \tag{464}$$

For a spatially local multicomponent system with a uniform temperature T, the inequality (463) assumes the form of the isothermal diffusional stability condition

$$-\frac{1}{T}\sum_{j,k}\int_{\mathcal{D}} dV\,\delta n_j(\vec{r},t)\left[\frac{\delta\mu_j(\vec{r},t)}{\delta n_k(\vec{r},t)}\right]\delta n_k(\vec{r},t) < 0 \tag{465}$$

postulated by Prigogine in nonequilibrium thermodynamics. In the above, $n_j(\vec{r},t)$ denotes the particle number density of species j with the chemical potential $\mu_j(\vec{r},t)$.

The first variation $\delta S(t)$ in the entropy, given by Eq. (458), vanishes only when the fluctuations $\delta\langle\hat{O}_j(\vec{r},t)\rangle$ are about the state of maximum realizable entropy for all conceivable values $\{\langle\hat{O}_j(\vec{r},t)\rangle_c\}$ of the thermodynamic coordinates $\{\langle\hat{O}_j(\vec{r},t)\rangle\}$, i.e., the state of equilibrium. This may or may not be the state described by $\hat{\bar{\rho}}(t)$. In general, $\hat{\bar{\rho}}(t)$ is the state of maximum entropy only for a given set of values $\{\langle\hat{O}_j(\vec{r},t)\rangle_v\}$ of the thermodynamic coordinates $\{\langle\hat{O}_j(\vec{r},t)\rangle\}$. For nonequilibrium systems, the first variation $\delta S(t)$ in the entropy does not vanish.

The aforementioned conditions for a system in equilibrium or nonequilibrium can be expressed as follows:

$$\begin{aligned} \delta S(t) &= 0 \qquad \text{equilibrium} \\ \delta S(t) &\neq 0 \qquad \text{nonequilibrium} \end{aligned} \tag{466}$$

or

$$k_{\mathrm{B}} \sum_j \int_{\mathcal{D}} dV\,\Lambda_j(\vec{r},t)\,\delta\langle\hat{O}_j(\vec{r},t)\rangle = 0 \quad \text{equilibrium}$$

$$k_{\mathrm{B}} \sum_j \int_{\mathcal{D}} dV\,\Lambda_j(\vec{r},t)\,\delta\langle\hat{O}_j(\vec{r},t)\rangle \neq 0 \quad \text{nonequilibrium} \tag{467}$$

If the thermodynamic coordinates $\langle\hat{O}_j(\vec{r},t)\rangle$ represent local densities $\{\rho_j(\vec{r},t)\}$ of conserved quantities, the equilibrium condition given by (467) implies that spatially uniform values of the thermodynamic parameters $\{\Lambda_j(\vec{r},t)\}$ represent a state of equilibrium. This result is in agreement with nonequilibrium thermodynamics.

Making use of the local entropy density from nonequilibrium thermodynamics, Prigogine formulated the stability condition

$$\frac{1}{2}\frac{d}{dt}\delta^2 S(t) = \sum_j \int_{\mathcal{D}} dV\, \delta\vec{X}_j(\vec{r},t)\cdot\delta\vec{J}_j(\vec{r},t) > 0 \tag{468}$$

for a spatially local system subject to energy and particle number density fluctuations. One can demonstrate that this stability condition follows from Eq. (459b) provided: (i) The matrix elements $\chi_{jk}(\vec{r},\vec{r}';t)$ of the susceptibility matrix $\boldsymbol{\chi}(t)$ are time independent. (ii) Fluctuations on the boundary between the system and its surroundings can be neglected. [The contribution of such fluctuations to the right side of the equality in (468) is given by $-k_{\rm B}\int_{\mathcal{S}} dS\,\delta\Lambda_j(\vec{r},t)\cdot\hat{n}\cdot\delta\vec{J}_j(\vec{r},t)$.] (iii) The requirements of a Lyapounov function are imposed upon $-\delta^2 S(t)$. Since $\hat{\bar{\rho}}(t)$ is time independent for a stationary state, the matrix elements $\chi_{jk}(\vec{r},\vec{r}';t)$ of the susceptibility matrix $\boldsymbol{\chi}(t)$ are time independent. With the exception of systems exhibiting small fluctuations about a stationary state, $-\delta^2 S(t)$ cannot, in general, be treated as a Lyapounov function.

VI. NONEQUILIBRIUM PROCESSES AND MICROSCOPIC DYNAMICS

A. QSM Theory

The quantum statistical mechanical theory of relaxation phenomena (QSM theory) is a maximum entropy approach to nonequilibrium processes that is in the same spirit as the statistical mechanical formulations of equilibrium and nonequilibrium thermodynamics given in Sections 3.3 and 5.4. As in the statistical mechanical formulation of nonequilibrium thermodynamics, the maximum entropy principle is assumed to apply to systems out of equilibrium.

QSM theory provides a firm statistical mechanical basis for the phenomenological theory of irreversible processes as formulated by Onsager and the version of classical Brownian motion discussed in Section 5.2. Moreover, it gives a number of formulas that can be employed in the investigation of the role of microscopic interactions in a diversity of nonequilibrium phenomena.

The roots of QSM theory lie in Mori's statistical mechanical theory of transport processes and Kubo's theory of thermal disturbances. The version of QSM theory given here with its refinements and modern embellishments was used by the author to develop an irreversible thermodynamic theory for photophysical phenomena and a quantum stochastic Fokker–Planck theory for adiabatic and nonadiabatic processes in condensed phases.

In QSM theory, we consider a composite system made up of the system of interest and its surroundings. In general, the system of interest and the surroundings do not occupy different spatial domains. Suppose, for example, we are interested in the excited dynamics of a collection of atoms/molecules enclosed in a vessel which is in thermal contact with a temperature reservoir. The surroundings include not only the temperature reservoir, but also the free electromagnetic field inside the vessel.

The macroscopic state of the composite system is defined by specifying the values of certain averaged quantities. The averaged quantities may or may not depend on position and/or time.

For the sake of simplicity, we confine our attention to composite systems characterized by either spatially independent averages $\{\langle\hat{O}_j(t)\rangle\}$ or spatially dependent averages $\{\langle\hat{O}_j(\vec{r},t)\rangle\}$. In general, both types of averages are required to specify the macroscopic state of the composite system.

Independent of the nature of the averages, we assume that the maximum entropy principle can be used to construct the statistical density operator $\hat{\bar{\rho}}(t)$ characterizing the macroscopic state of the composite system at time t. As before, we require the averages to be linearly independent.

1. Spatially Independent Properties

Consider a composite system made up of the system of interest and its surroundings. The macroscopic state of the composite system is defined by its average energy $\langle\hat{\mathcal{H}}\rangle$ and some unspecified averages $\{\langle\hat{O}_j(t)\rangle\}$. The average energy $\langle\hat{\mathcal{H}}\rangle$ is assumed to be time independent. Usually, we take the unspecified averages $\{\langle\hat{O}_j(t)\rangle\}$ to be solely associated with the system of interest.

Adopting the time-dependent version of the maximum entropy principle, we find that the statistical density operator $\hat{\bar{\rho}}(t)$ describing the macroscopic state of the composite system is given by

$$\hat{\bar{\rho}}(t) = \exp\left[\Omega(t)\hat{I} - \beta\hat{\mathcal{H}} - \sum_j \Lambda_j(t)\hat{O}_j\right], \tag{469}$$

where

$$\Omega(t) = -\ln Z(t), \tag{470}$$

with

$$Z(t) = \mathrm{Tr}\exp\left[-\beta\hat{\mathcal{H}} - \sum_j \Lambda_j(t)\hat{O}_j\right]. \tag{471}$$

In view of Eq. (340), we can write the entropy $S(t)$ of the macroscopic state described by $\hat{\bar{\rho}}(t)$ as

$$S(t) = -k_{\rm B}\Omega(t) + k_{\rm B}\beta\langle\hat{\mathcal{H}}\rangle + k_{\rm B}\sum_j \Lambda_j(t)\langle\hat{O}_j(t)\rangle. \tag{472}$$

Making use of this relation, one can demonstrate that

$$dS(t) = k_B\beta d\langle\hat{\mathcal{H}}\rangle + k_B \sum_j \Lambda_j(t)\, d\langle\hat{O}_j(t)\rangle. \quad (473)$$

It follows from Eq. (473) that the partial derivatives of the entropy $S(t)$ are given by

$$\frac{\partial S(t)}{\partial\langle\hat{\mathcal{H}}\rangle} = k_B\beta \quad (474)$$

and

$$\frac{\partial S(t)}{\partial\langle\hat{O}_j(t)\rangle} = k_B\Lambda_j(t). \quad (475)$$

Making the identification $X_j(t) = k_B\Lambda_j(t)$, we find that the latter equation is equivalent to Onsager's definition of the thermodynamic driving force $X_j(t)$ in the phenomenological theory of irreversible processes for the case of observables with a discrete index (see Section V.C.3).

Let us consider the response of the system to the displacements $\langle\delta\hat{O}_j(t)\rangle = \langle\hat{O}_j(t)\rangle - \langle\hat{O}_j\rangle_{eq}$ from thermal equilibrium, where $\langle\hat{O}_j\rangle_{eq} = \mathrm{Tr}\,\hat{\rho}_{eq}\hat{O}_j$, with $\hat{\rho}_{eq} = \exp(\Omega\hat{I} - \beta\hat{\mathcal{H}})$. We define the observed response in terms of the phenomenological time derivatives

$$\frac{d}{dt}\langle\delta\hat{O}_j(t;\Delta t)\rangle = \frac{\langle\delta\hat{O}_j(t+\Delta t)\rangle - \langle\delta\hat{O}_j(t)\rangle}{\Delta t} \quad (476a)$$

$$= \Delta t^{-1}\int_t^{t+\Delta t} dt' \frac{d}{dt'}\langle\delta\hat{O}_j(t')\rangle. \quad (476b)$$

As in classical Brownian motion theory, the phenomenological time derivative is a coarse-grained time derivative obtained by time averaging the instantaneous time derivative $(d/dt')\langle\delta\hat{O}_j(t')\rangle$ over the time scale Δt of macroscopic measurements (time resolution of the macroscopic observer).

In Eq. (476a), the quantity $\langle\delta\hat{O}_j(t+\Delta t)\rangle$ denotes a conditional average defined by

$$\langle\delta\hat{O}_j(t+\Delta t)\rangle = \langle\delta\hat{O}_j(t+\Delta t \mid \{\langle\hat{O}_k(t)\rangle\})\rangle \quad (477a)$$

$$= \mathrm{Tr}\,\hat{\rho}(t)\delta\hat{O}_j(\Delta t), \quad (477b)$$

where $\delta\hat{O}_j(\Delta t) = \exp(+i\hat{\mathcal{L}}\Delta t)\delta\hat{O}_j$, with $\delta\hat{O}_j = \hat{O}_j - \langle\hat{O}_j\rangle_{eq}$. The conditional average is defined in such a way that $\langle\delta\hat{O}_j(t+\Delta t)\rangle$ represents the average value of the displacement $\delta\hat{O}_j$ from equilibrium at time $t+\Delta t$ given that the averages possess the values $\{\langle\hat{O}_k(t)\rangle\}$ at time t.

The conditional average $\langle\hat{O}_j(t+\Delta t)\rangle = \langle\hat{O}_j(t+\Delta t\,|\{\langle\hat{O}_k(t)\rangle\})\rangle$ is formally equivalent to the average $\overline{O_j(t+\Delta t\,|\,\mathbf{O}')}$ appearing in Onsager's treatment of irreversible processes. In view of this equivalence and the equality

$$\langle\delta\hat{O}_j(t+\Delta t)\rangle - \langle\delta\hat{O}_j(t)\rangle = \langle\hat{O}_j(t+\Delta t)\rangle - \langle\hat{O}_j(t)\rangle$$

we find that Onsager's definition of the phenomenological time derivative $\overline{dO_j(t+\Delta t|O')}/dt$ is equivalent to Eq. (476a). [See Eq. (338).]

We can recast Eq. (476a) in the form

$$\frac{d}{dt}\langle\delta\hat{O}_j(t;\Delta t)\rangle = \langle\delta\hat{O}_j(t)\rangle + \int_0^{\Delta t} d\tau(1-\tau/\Delta t)\langle\delta\hat{\dot{O}}_j(t+\tau)\rangle, \quad (478)$$

where $\langle\delta\hat{O}_j(t)\rangle$ and $\langle\delta\hat{\dot{O}}_j(t+\tau)\rangle$ are the conditional averages

$$\langle\delta\hat{O}_j(t)\rangle = \langle\delta\hat{O}_j(t \mid \{\langle\hat{O}_k(t)\rangle\})\rangle \quad (479a)$$

$$= \mathrm{Tr}\,\hat{\bar{\rho}}(t)\,\delta\hat{O}_j(0) \quad (479b)$$

and

$$\langle\delta\hat{\dot{O}}_j(t+\tau)\rangle = \langle\delta\hat{\dot{O}}_j(t+\tau \mid \{\langle\hat{O}_k(t)\rangle\})\rangle \quad (480a)$$

$$= \mathrm{Tr}\,\hat{\bar{\rho}}(t)\delta\hat{\dot{O}}_j(\tau). \quad (480b)$$

In order to determine suitable expressions for the quantities $(d/dt)\langle\delta\hat{O}_j(t;\Delta t)\rangle$ and $\langle\delta\hat{O}_j(t)\rangle$, we find it convenient to rewrite Eq. (469) as

$$\hat{\bar{\rho}}(t) = \exp[\Delta\Omega(t)]\,\hat{\rho}_{eq}\left[\hat{I} - \beta^{-1}\sum_k \Lambda_k(t) \times \int_0^\beta d\lambda\hat{U}(i\hbar\lambda)\hat{O}_k(0)\hat{\mathcal{R}}(\lambda,t)\right], \quad (481)$$

where $\Delta\Omega(t) = \Omega(t) - \Omega$, $\hat{U}(i\hbar\lambda) = \exp(\lambda\hat{\mathcal{H}})$, and $\hat{\mathcal{R}}(\lambda,t) = \exp\{-\lambda[\hat{\mathcal{H}} + \beta^{-1}\sum_j \Lambda_j(t)\hat{O}_j]\}$.

Making use of Eq. (481) to evaluate the formal expressions for $(d/dt)\langle\delta\hat{O}_j(t;\Delta t)\rangle$ and $\langle\delta\hat{O}_j(t)\rangle$, we obtain

$$\frac{d}{dt}\langle\delta\hat{O}_j(t;\Delta t)\rangle = -\exp[\Delta\Omega(t)]\sum_k \Lambda_k(t) \times \left\{\beta^{-1}\int_0^\beta d\lambda\langle\hat{U}(i\hbar\lambda)\,\hat{O}_k(0)\,\hat{\mathcal{R}}(\lambda,t)\,\delta\hat{O}_j(0)\rangle_{eq} + \beta^{-1}\int_0^{\Delta t} d\tau\,(1-\tau/\Delta t) \times \int_0^\beta d\lambda\langle\hat{U}(i\hbar\lambda)\hat{O}_k(0)\,\hat{\mathcal{R}}(\lambda,t)\,\delta\hat{\dot{O}}_j(\tau)\rangle_{eq}\right\} \quad (482)$$

and

$$\langle\delta\hat{O}_j(t)\rangle = -\exp[\Delta\Omega(t)]\sum_k \Lambda_k(t)\beta^{-1}\int_0^\beta d\lambda\langle\hat{U}(i\hbar\lambda) \times \hat{O}_k(0)\hat{\mathcal{R}}(\lambda,t)\delta\hat{O}_j(0)\rangle_{eq}. \quad (483)$$

The thermodynamic parameters $\{\Lambda_j(t)\}$ can be obtained by solving Eq. (483). Given these quantities, the response to the displacements from equilibrium can be determined using Eq. (482).

Both $(d/dt)\langle\delta\hat{O}_j(t;\Delta t)\rangle$ and $\langle\delta\hat{O}_j(t)\rangle$ are nonlinear functions of the thermodynamic parameters $\{\Lambda_j(t)\}$. If the system is sufficiently close to thermal equilibrium, these quantities should be well approximated by the following linearized forms of Eqs. (482) and (483):

$$\frac{d}{dt}\langle\delta\hat{O}_j(t;\Delta t)\rangle = \sum_k M_{jk}\Lambda_k(t) \tag{484}$$

and

$$\langle\delta\hat{O}_j(t)\rangle = -\sum_k \chi_{jk}\Lambda_k(t), \tag{485}$$

where

$$M_{jk} = M^S_{jk} + M^C_{jk} \tag{486}$$

and

$$\chi_{jk} = \beta^{-1}\int_0^\beta \langle\delta\hat{O}_k(-i\hbar\lambda)\delta\hat{O}_j(0)\rangle_{\rm eq}, \tag{487}$$

with

$$M^S_{jk} = -\beta^{-1}\int_0^\beta d\lambda\langle\delta\hat{O}_k(-i\hbar\lambda)\delta\dot{\hat{O}}_j(0)\rangle_{\rm eq} \tag{488}$$

and

$$M^C_{jk} = \beta^{-1}\int_0^{\Delta t} dt(1-t/\Delta t) \times\int_0^\beta d\lambda\langle\delta\hat{\dot{O}}_k(-t-i\hbar\lambda)\delta\hat{\dot{O}}_j(0)\rangle_{\rm eq}. \tag{489}$$

In the classical limit, we can write

$$M_{jk} = -\langle\delta O_k(0)\delta\dot{O}_j(0)\rangle_{\rm eq} + \int_0^{\Delta t} dt(1-t/\Delta t)\langle\delta\dot{O}_k(-t)\delta\dot{O}_j(0)\rangle_{\rm eq} \tag{490}$$

and

$$\chi_{jk} = \langle\delta O_k(0)\delta O_j(0)\rangle_{\rm eq}, \tag{491}$$

where $\delta O_j(0) = O_j - \langle O_j\rangle_{\rm eq}$ and $\delta O_k(0) = O_k - \langle O_k\rangle_{\rm eq}$ are classical dynamical variables, with $\langle\,\rangle_{\rm eq}$ denoting an average over the classical equilibrium canonical probability density for the composite system.

Making the identifications $L_{jk} = M_{jk}/k_{\rm B}$ and $X_k(t) = k_{\rm B}\Lambda_j(t)$, we find that Eq. (484) is identical in form to the linear equations in Onsager's phenomenological theory of irreversible processes for the case of observables with a discrete index. [See Eqs. (337) and (475).] If the average $\overline{O_j(t+\Delta t)O_k(t)}$ in Onsager's expression for the phenomenological coefficient L_{jk} [see Eq. (339)] is interpreted as an average over the classical equilibrium canonical probability density for the composite system and the average is assumed to be stationary, i.e., $\overline{O_j(t+\Delta t)O_k(t)} = \overline{O_j(\Delta t)O_k(0)}$, one can demonstrate that Onsager's expression for L_{jk} is equivalent to the classical result for $M_{jk}/k_{\rm B}$. In view of these considerations and our earlier observations, we conclude that the linear phenomenological equations adopted by Onsager in the treatment of irreversible processes correspond to the classical limit of Eq. (484).

The results given by Eqs. (484) and (485) enable us to construct the following linear equation of motion for the displacements from equilibrium:

$$\frac{d}{dt}\langle\delta\hat{O}_j(t;\Delta t)\rangle = -\sum_k[i\Omega_{jk} + K_{jk}]\langle\delta\hat{O}_k(t)\rangle, \tag{492}$$

where

$$i\Omega_{jk} = \sum_l M^S_{jl}\chi^{-1}_{lk} \tag{493}$$

and

$$K_{jk} = \sum_l M^C_{jl}\chi^{-1}_{lk}, \tag{494}$$

with χ^{-1}_{lk} denoting a matrix element of the inverse $\boldsymbol{\chi}^{-1}$ of the matrix $\boldsymbol{\chi}$.

From a macroscopic point of view, Eq. (492) can be treated as a differential equation. Treated as such, the displacements from equilibrium $\{\langle\delta\hat{O}_j(t)\rangle\}$ will exhibit damped oscillatory temporal behavior and decay to zero in the infinite-time limit provided all of the eigenvalues of the matrix $i\boldsymbol{\Omega} + \mathbf{K}$ have positive real parts.

As in classical Brownian motion theory, the damping in Eq. (492), characterized by the damping constants $\{K_{jk}\}$, arises from spontaneous equilibrium fluctuations described by equilibrium time correlation functions. The oscillatory motion of the system is characterized by the frequencies $\{\Omega_{jk}\}$. The influence of microscopic interactions on the overall time evolution of the displacements from equilibrium is buried in the quantities $\{\Omega_{jk}\}$ and $\{K_{jk}\}$ or equivalently $\{M^S_{jk}\}$, $\{M^C_{jk}\}$, and $\{\chi_{jk}\}$.

In the actual evaluation of the phenomenological coefficients $\{M^C_{jk}\}$, the integration limit Δt is usually extended to $+\infty$ and the time correlation function is multiplied by the convergence factor $\exp(-\epsilon t)$, where $\epsilon \to 0^+$ at the end of the calculation. More specifically, M^C_{jk} is written as

$$M^C_{jk} = \lim_{\epsilon\to 0^+}\beta^{-1}\int_0^\infty dt\exp(-\epsilon t) \times\int_0^\beta d\lambda\langle\delta\hat{O}_k(-t-i\hbar\lambda)\delta\dot{\hat{O}}_j(0)\rangle_{\rm eq}. \tag{495}$$

The formal expressions Eqs. (487)–(489) for χ_{jk}, M^S_{jk}, and M^C_{jk} can be given matrix representations similar to the matrix representation $\langle O^\dagger|\rho_t\rangle$ of the average $\langle\hat{O}(t)\rangle$:

$$\chi_{jk} = \langle \delta O_k^\dagger | \delta\tilde{O}_j \rho_{\rm eq}\rangle, \tag{496}$$

$$M^S_{jk} = \begin{cases} -\langle \delta O_k^\dagger | \delta\tilde{O}_j \rho_{\rm eq}\rangle \\ -\langle \delta \dot{O}_j^\dagger | \delta\tilde{O}_k \rho_{\rm eq}\rangle \\ \langle \delta O_j^\dagger | \delta\tilde{O}_k \rho_{\rm eq}\rangle, \end{cases} \tag{497}$$

and

$$M^C_{jk} = \int_0^{\Delta t} dt(1 - t/\Delta t)\langle \delta \dot{O}_j^\dagger | \exp(-i\hat{\mathcal{L}}t)|\delta\tilde{\tilde{O}}_k\rho_{\rm eq}\rangle. \tag{498}$$

In order to simplify the notation, we have made use of the definition $\hat{\tilde{A}} = \beta^{-1}\int_0^\beta d\lambda\, \hat{A}(i\hbar\lambda)$ of the Kubo transform $\hat{\tilde{A}}$ of the operator $\hat{A}$.

The approximate form given by Eq. (495) for M^C_{jk} can be written

$$M^C_{jk} = \lim_{\epsilon\to 0^+}\int_0^\infty dt\exp(-\epsilon t)\langle \delta \dot{O}_j^\dagger | \exp(-i\hat{\mathcal{L}}t)|\delta\tilde{O}_k\rho_{\rm eq}\rangle. \tag{499}$$

2. Spatially Dependent Properties

Consider a composite system made up of the system of interest and its surroundings. The macroscopic state of the composite system is defined by the spatially dependent averages $\{\langle\hat{O}_j(\vec{r},t)\rangle\}$. For the sake of simplicity, we assume that these averages are the average local densities $\{\langle\hat{\rho}_j(\vec{r},t)\rangle\}$ of conserved quantities. We take $\langle\hat{\rho}_1(\vec{r},t)\rangle$ to represent the average local energy density of the composite system at the space-time point $(\vec{r},t)$.

Adopting the time-dependent version of the maximum entropy principle, we find that the statistical density operator $\hat{\tilde{\rho}}(t)$ describing the macroscopic state of the composite system is given by

$$\hat{\tilde{\rho}}(t) = Z(t)^{-1}\exp\left[-\sum_j\int_{\mathcal{D}} dV\,\Lambda_j(\vec{r},t)\hat{\rho}_j(\vec{r})\right], \tag{500}$$

where

$$Z(t) = \mathrm{Tr}\,\exp\left[-\sum_j\int_{\mathcal{D}} dV\,\Lambda_j(\vec{r},t)\hat{\rho}_j(\vec{r})\right]. \tag{501}$$

The volume integrations in Eqs. (500) and (501) extend over the spatial domain $\mathcal{D}$ occupied by the composite system.

It is convenient to write the thermodynamic parameters $\Lambda_j(\vec{r},t)$ as $\Lambda_j(\vec{r},t) = \Lambda_j^{(0)} + \Lambda_j^{(1)}(\vec{r},t)$, where $\Lambda_j^{(0)}$ is the value assumed by $\Lambda_j(\vec{r},t)$ when the composite system is in the state of equilibrium. Assuming that the state of equilibrium is the state of thermal equilibrium described by the equilibrium canonical density operator $\hat{\rho}_{\rm eq} = Z^{-1}\exp(-\beta\hat{\mathcal{H}})$, where $\hat{\mathcal{H}}$ is the Hamiltonian for the composite system, we must have $\Lambda_1^{(0)} = \beta$ and $\Lambda_j^{(0)} = 0$ for $j \neq 1$. In order to simplify the notation, we write $\Lambda_1(\vec{r},t) = \beta + \lambda_1(\vec{r},t)$ and $\Lambda_j(\vec{r},t) = \lambda_j(\vec{r},t)$ for $j\neq 1$, where $\lambda_j(\vec{r},t) = \Lambda_j^{(1)}(\vec{r},t)$ for all j.

Making the substitutions $\Lambda_1(\vec{r},t) = \beta + \lambda_1(\vec{r},t)$ and $\Lambda_j(\vec{r},t) = \lambda_j(\vec{r},t)$ for $j\neq 1$ in Eqs. (500) and (501), we obtain

$$\hat{\tilde{\rho}}(t) = Z(t)^{-1}\exp\left[-\beta\hat{\mathcal{H}} - \sum_j\int_{\mathcal{D}} dV\lambda_j(\vec{r},t)\hat{\rho}_j(\vec{r})\right], \tag{502}$$

where

$$Z(t) = \mathrm{Tr}\,\exp\left[-\beta\hat{\mathcal{H}} - \sum_j\int_{\mathcal{D}} dV\lambda_j(\vec{r},t)\hat{\rho}_j(\vec{r})\right]. \tag{503}$$

Adopting an approach similar to the one adopted in the treatment of spatially independent thermodynamic coordinates, we find that the response of the system for sufficiently small displacements from thermal equilibrium is given by

$$\frac{\partial}{\partial t}\langle\hat{\rho}_j(\vec{r},t;\Delta t)\rangle = \sum_k\int_{\mathcal{D}} dV' M_{jk}(\vec{r},\vec{r}')\lambda_k(\vec{r}',t), \tag{504}$$

where

$$M_{jk}(\vec{r},\vec{r}') = M^S_{jk}(\vec{r},\vec{r}') + M^C_{jk}(\vec{r},\vec{r}'), \tag{505}$$

with

$$M^S_{jk}(\vec{r},\vec{r}') = \begin{cases} -\langle \rho_k^\dagger(\vec{r}')|\tilde{\rho}_j(\vec{r})\rho_{\rm eq}\rangle \\ -\langle \dot{\rho}_j^\dagger(\vec{r})|\tilde{\rho}_k(\vec{r}')\rho_{\rm eq}\rangle \\ \langle \rho_j^\dagger(\vec{r})|\tilde{\rho}_k(\vec{r}')\rho_{\rm eq}\rangle \end{cases} \tag{506}$$

and

$$M^C_{jk}(\vec{r},\vec{r}') = \int_0^{\Delta t} dt(1 - t/\Delta t)\langle\dot{\rho}_j^\dagger(\vec{r})| \times \exp(-i\hat{\mathcal{L}}t)|\tilde{\tilde{\rho}}_k(\vec{r}')\rho_{\rm eq}\rangle. \tag{507}$$

Since the operators $\{\hat{\rho}_j(\vec{r})\}$ represent local densities of conserved quantities, we can write $\int_{\mathcal{D}} dV\,\hat{\rho}_j(\vec{r}) = \hat{0}$, where $\hat{0}$ is the null operator. Making use of this relation and Eqs. (506) and (507), we find that $\int_{\mathcal{D}} dV'\vec{M}_{jk}(\vec{r},\vec{r}') = 0$. In view of this identity and the relations $\Lambda_1(\vec{r},t) = \beta + \lambda_1(\vec{r},t)$ and $\Lambda_k(\vec{r},t) = \lambda_k(\vec{r},t)$ for $k\neq 1$, we can rewrite Eq. (504) as

$$\frac{\partial}{\partial t}\langle\hat{\rho}_j(\vec{r},t;\Delta t)\rangle = \sum_k\int_{\mathcal{D}} dV' M_{jk}(\vec{r},\vec{r}')\Lambda_k(\vec{r}',t), \tag{508}$$

where the $\{\Lambda_k(\vec{r}', t)\}$ are the original thermodynamic parameters.

Making use of the microscopic continuity equation $\hat{\rho}_j(\vec{r}) = -\vec{\nabla}\cdot\hat{\vec{J}}_j(\vec{r})$, where $\hat{\vec{J}}_j(\vec{r})$ is a current density operator, we find that Eq. (508) can be cast in the form

$$\langle \hat{\vec{J}}_j(\vec{r}, t; \Delta t)\rangle = \sum_k \left[\int_{\mathcal{D}} dV' \vec{\mathfrak{M}}^S_{jk}(\vec{r}, \vec{r}')\Lambda_k(\vec{r}', t) + \int_{\mathcal{D}} dV' \vec{\vec{\mathfrak{M}}}^C_{jk}(\vec{r}, \vec{r}')\cdot\vec{\nabla}'\Lambda_k(\vec{r}', t)\right], \tag{509}$$

where

$$\vec{\mathfrak{M}}^S_{jk}(\vec{r}, \vec{r}') = \begin{cases} -\langle \vec{J}^\dagger_j(\vec{r})\big|\tilde{O}_k(\vec{r}')\rho_{\text{eq}}\rangle \\ -\langle O^\dagger_k(\vec{r}')\big|\vec{J}_j(\vec{r})\rho_{\text{eq}}\rangle \end{cases} \tag{510}$$

and

$$\vec{\vec{\mathfrak{M}}}^C_{jk}(\vec{r}, \vec{r}') = \int_0^{\Delta t} dt\,(1 - t/\Delta t)\langle \vec{J}^\dagger_j(\vec{r})\big| \times \exp(-i\hat{\mathcal{L}}t)\big|\vec{J}_k(\vec{r}')\rho_{\text{eq}}\rangle. \tag{511}$$

Equation (509) represents a spatially nonlocal generalization of the linear phenomenological equations in nonequilibrium thermodynamics for the case of continuous systems. [See Eq. (320).] As in nonequilibrium thermodynamics, the gradients $k_B\vec{\nabla}\Lambda_j(\vec{r}, t)$ of the spatially dependent thermodynamic parameters $k_B\Lambda_j(\vec{r}, t)$, where $k_B\Lambda_j(\vec{r}, t) = \delta S(t)/\delta\langle\hat{\rho}_j(\vec{r}, t\rangle$, can be treated as driving forces $\vec{X}_k(\vec{r}, t)$ for the currents $\langle\hat{\vec{J}}_j(\vec{r}, t; \Delta t)\rangle$. Of course, one must interpret the quantity $\vec{J}_j(\vec{r}, t)$ appearing in the linear phenomenological equations as $\langle\hat{\vec{J}}_j(\vec{r}, t; \Delta t)\rangle$.

Similar to the approximation of Eq. (498) by Eq. (499), the formal expression given by Eq. (511) for $\vec{\vec{\mathfrak{M}}}^C_{jk}(\vec{r}, \vec{r}')$ is sometimes written

$$\vec{\vec{\mathfrak{M}}}^C_{jk}(\vec{r}, \vec{r}') = \lim_{\epsilon\to 0^+}\int_0^\infty dt\exp(-\epsilon t)\langle \vec{J}^\dagger_j(\vec{r})\big| \times \exp(-i\hat{\mathcal{L}}t)\big|\vec{J}_k(\vec{r}')\rho_{\text{eq}}\rangle. \tag{512}$$

B. Response to External Disturbances

Let us consider the problem of describing the response of a system to some probe, such as an electric or magnetic field, that can be treated as a classical external force. This problem can be dealt with by starting with the following version of the quantum Liouville equation:

$$\frac{d}{dt}\hat{\rho}(t) = -i\hat{\hat{\mathcal{L}}}(t)\hat{\rho}(t), \tag{513}$$

where $\hat{\hat{\mathcal{L}}}(t)$ is a time-dependent Liouville operator given by $\hat{\hat{\mathcal{L}}}(t) = \hat{\hat{\mathcal{L}}} + \hat{\hat{\mathcal{L}}}_{\text{ext}}(t)$.

The Liouville operator $\hat{\hat{\mathcal{L}}} = (1/\hbar)\hat{\hat{\mathcal{H}}}^-$ embodies the dynamics of the system in the absence of the external force, where the superoperator $\hat{\hat{\mathcal{H}}}^-$ is formed using the system Hamiltonian $\hat{\mathcal{H}}$. The influence of the external force on the system is described by $\hat{\hat{\mathcal{L}}}_{\text{ext}}(t) = (1/\hbar)\hat{\hat{\mathcal{H}}}^-_{\text{ext}}(t)$, where the superoperator $\hat{\hat{\mathcal{H}}}^-_{\text{ext}}(t)$ is formed from the interaction Hamiltonian $\hat{\mathcal{H}}_{\text{ext}}(t)$ describing the coupling between the external force and the system.

Consider, for example, the interaction $\hat{\mathcal{H}}_{\text{ext}}(t)$ between a polarizable medium and the classical electric field $\vec{E}(\vec{r}, t)$. For this case, we write

$$\hat{\mathcal{H}}_{\text{ext}}(t) = -\int_{\mathcal{D}} dV\vec{E}(\vec{r}, t)\cdot\hat{\vec{P}}(\vec{r}), \tag{514}$$

where the integration is over the spatial domain $\mathcal{D}$ occupied by the system and its surroundings and

$$\hat{\vec{P}}(\vec{r}) = \frac{1}{2}\sum_m [\hat{\vec{\mu}}_m, \delta(\vec{r} - \hat{\vec{q}}_m)]_+ \tag{515}$$

is the electric polarization operator for the system at the position $\vec{r}$, with the subscript $+$ indicating that the quantity $[\hat{\vec{\mu}}_m, \delta(\vec{r} - \hat{\vec{q}}_m)]_+$ is an anticommutator. The index m in Eq. (515) runs over the constituent atoms/molecules with $\hat{\vec{\mu}}_m$ and $\hat{\vec{q}}_m$ denoting the dipole moment and coordinate operators, respectively, for atom/molecule m.

The interaction Hamiltonian $\hat{\mathcal{H}}_{\text{ext}}(t)$ given by Eq. (514) describes the coupling of a classical force $\vec{F}_{\hat{\vec{B}}}(\vec{r}, t) = \vec{E}(\vec{r}, t)$ with a local vector property $\hat{\vec{B}}(\vec{r}) = \hat{\vec{P}}(\vec{r})$ of the system. Restricting our considerations to systems involving this type of coupling, we assume that $\hat{\mathcal{H}}_{\text{ext}}(t)$ can be written in the form

$$\hat{\mathcal{H}}_{\text{ext}}(t) = -\int_{\mathcal{D}} dV\vec{F}_{\hat{\vec{B}}}(\vec{r}, t)\cdot\hat{\vec{B}}(\vec{r}), \tag{516}$$

where

$$\hat{\vec{B}}(\vec{r}) = \frac{1}{2}\sum_m [\hat{\vec{B}}_m, \delta(\vec{r} - \hat{\vec{q}}_m)]_+. \tag{517}$$

For this model,

$$\hat{\hat{\mathcal{L}}}_{\text{ext}}(t) = -\frac{1}{\hbar}\int_{\mathcal{D}} dV\vec{F}_{\hat{\vec{B}}}(\vec{r}, t)\cdot\hat{\hat{\vec{B}}}^-(\vec{r}). \tag{518}$$

It is convenient to recast the Liouville equation given by Eq. (513) in the vector form

$$\frac{d}{dt}|\rho_t\rangle = -\hat{L}(t)|\rho_t\rangle, \tag{519}$$

where the transition operator $\hat{L}(t)$ is given by $\hat{L}(t) = i\hat{\mathcal{L}}(t)$. The formal solution $|\rho_t\rangle$ of this equation can be expressed in terms of the integral equation

$$|\rho_t\rangle = \exp[-\hat{L}(t-t_0)]|\rho_{t_0}\rangle - \int_{t_0}^{t} dt' \exp[-\hat{L}(t-t')]\,\hat{L}_{\text{ext}}(t')|\rho_{t'}\rangle, \quad (520)$$

with $\hat{L} = i\hat{\mathcal{L}}$ denoting the transition operator of the system and $|\rho_{t_0}\rangle$ representing its initial state.

Assuming the system is initially prepared in the state of thermal equilibrium, we write $|\rho_{t_0}\rangle = |\rho_{\text{eq}}\rangle$, where $|\rho_{\text{eq}}\rangle$ is a vector corresponding to the equilibrium canonical density operator. Introducing this assumption, we obtain for Eq. (520)

$$|\rho_t\rangle = |\rho_{\text{eq}}\rangle - \int_{t_0}^{t} dt' \exp[-\hat{L}(t-t')\,]\,\hat{L}_{\text{ext}}(t')|\rho_{t'}\rangle. \quad (521)$$

In view of the formal structure of $\hat{\hat{L}}_{\text{ext}}(t) = i\hat{\hat{\mathcal{L}}}_{\text{ext}}(t)$, we can write

$$\hat{L}_{\text{ext}}(t')|C\rangle = -\frac{i}{\hbar}\int_{\mathcal{D}} dV' |[\vec{B}(\vec{r}',0), C]_-\rangle \cdot \vec{F}_{\vec{B}}(\vec{r}',t') \quad (522)$$

for any vector $|C\rangle$. Making use of this relation, we find that Eq. (521) can be written

$$|\rho_t\rangle = |\rho_{\text{eq}}\rangle + \frac{i}{\hbar}\int_{t_0}^{t} dt' \int_{\mathcal{D}} dV' \times \exp[-\hat{L}(t-t')]|[\vec{B}(\vec{r}',0), \rho_{t'}]_-\rangle \cdot \vec{F}_{\vec{B}}(\vec{r}',t'). \quad (523)$$

With Eq. (523) at our disposal, we can proceed to consider the influence of the external force $\vec{F}_{\vec{B}}(\vec{r},t)$ on the properties of the system. Using Eq. (523) to form the inner product $\langle \vec{A}^{\dagger}(\vec{r})|\rho_t\rangle$ in the expression $\langle \hat{\vec{A}}(\vec{r},t)\rangle = \langle \vec{A}^{\dagger}(\vec{r})|\rho_t\rangle$, we obtain

$$\langle \delta\hat{\vec{A}}(\vec{r},t)\rangle = \int_{t_0}^{t} dt' \int_{\mathcal{D}} dV' \vec{\vec{\Phi}}_{\vec{A}\vec{B}}(\vec{r}\,t,\vec{r}'t') \cdot \vec{F}_{\vec{B}}(\vec{r}',t'), \quad (524)$$

where $\langle \delta\hat{\vec{A}}(\vec{r},t)\rangle = \langle \hat{\vec{A}}(\vec{r},t)\rangle - \langle \hat{\vec{A}}(\vec{r})\rangle_{\text{eq}}$ and

$$\vec{\vec{\Phi}}_{\vec{A},\vec{B}}(\vec{r}\,t,\vec{r}'\,t') = \frac{i}{\hbar}\big\langle \vec{A}^{\dagger}_{t-t'}(\vec{r})\,\big|\,[\vec{B}(\vec{r}',0), \rho_{t'}]_-\big\rangle. \quad (525)$$

Equation (524) is an exact result for the displacement $\langle \delta\hat{\vec{A}}(\vec{r},t)\rangle$ from equilibrium arising from the external force $\vec{F}_{\vec{B}}(\vec{r},t)$. As evident from Eq. (524), the response of the system to the external force $\vec{F}_{\vec{B}}(\vec{r},t)$ is nonlocal in space and time. Moreover, the response is nonlinear with the nonlinear dependence carried by the nonequilibrium time correlation function $\vec{\vec{\Phi}}_{\vec{A}\vec{B}}(\vec{r}t,\vec{r}'t')$ through $|\rho_{t'}\rangle$.

The result given by Eq. (525) can be rewritten as

$$\vec{\vec{\Phi}}_{\vec{A}\vec{B}}(\vec{r}\,t,\vec{r}'t') = \frac{i}{\hbar}\langle[\hat{\vec{A}}(\vec{r},t-t'),\,\hat{\vec{B}}(\vec{r}',0)]_-\rangle_{\rho(t')} \quad (526a)$$

$$= \frac{i}{\hbar}\,\text{Tr}\,\hat{\rho}(t')[\hat{\vec{A}}(\vec{r},t-t'),\,\hat{\vec{B}}(\vec{r}',0)]_-. \quad (526b)$$

Since $\vec{\vec{\Phi}}_{\vec{A}\vec{B}}(\vec{r}\,t,\vec{r}'t')$ represents the response of the system at the space-time point $(\vec{r},t)$ to the disturbance at the space-time point $(\vec{r}',t')$, we refer to $\vec{\vec{\Phi}}_{\vec{A}\vec{B}}(\vec{r}\,t,\vec{r}'t')$ as a response function.

Suppose $\vec{F}_{\vec{B}}(\vec{r},t)$ is a CW monochromatic force of the form

$$\vec{F}_{\vec{B}}(\vec{r},t) = \frac{1}{2}\sum_{\vec{k}}\big\{\vec{F}_{\vec{B}}(\vec{k},\omega)\exp[+i(\vec{k}\cdot\vec{r}-\omega t)] + \vec{F}^{*}_{\vec{B}}(\vec{k},\omega)\exp[-i(\vec{k}\cdot\vec{r}-\omega t)]\big\}. \quad (527)$$

Substituting this form into Eq. (524), we obtain

$$\langle \delta\hat{\vec{A}}(\vec{r},t)\rangle = \text{Re}\sum_{\vec{k}} \vec{\vec{\chi}}_{\vec{A}\vec{B}}(\vec{k},\omega)\cdot\vec{F}_{\vec{B}}(\vec{k},\omega) \times \exp[+i(\vec{k}\cdot\vec{r}-\omega t)], \quad (528)$$

where

$$\vec{\vec{\chi}}_{\vec{A}\vec{B}}(\vec{k},\omega) = \int_{t_0}^{t} dt' \exp[+i\omega(t-t')] \times \int_{\mathcal{D}} dV' \exp[+i\vec{k}\cdot(\vec{r}'-\vec{r})]\vec{\vec{\Phi}}_{\vec{A}\vec{B}}(\vec{r}t,\vec{r}'t'). \quad (529)$$

The wavevector- and frequency-dependent quantity $\vec{\vec{\chi}}_{\vec{A}\vec{B}}(\vec{k},\omega)$ can be regarded as a generalized nonlinear susceptibility characterizing the response of a system to a monochromatic force.

For a sufficiently weak external force $\vec{F}_{\vec{B}}(\vec{r},t)$, the departure $|\delta\rho_t\rangle = |\rho_t\rangle - |\rho_{\text{eq}}\rangle$ from equilibrium is expected to be small. Then $|\rho_t\rangle \approx |\rho_{\text{eq}}\rangle$. Assuming this to be the case, the nonequilibrium response function $\vec{\vec{\Phi}}_{\vec{A}\vec{B}}(\vec{r}\,t,\vec{r}'t')$ should be well approximated by the equilibrium response function

$$\vec{\vec{\Phi}}_{\vec{A}\vec{B}}(\vec{r},\vec{r}';t-t') = \frac{i}{\hbar}\langle[\hat{\vec{A}}(\vec{r},t-t'),\,\hat{\vec{B}}(\vec{r}',0)]_-\rangle_{\text{eq}} \quad (530a)$$

$$= \frac{i}{\hbar}\,\text{Tr}\,\hat{\rho}_{\text{eq}}[\hat{\vec{A}}(\vec{r},t-t'),\,\hat{\vec{B}}(\vec{r}',0)]_-. \quad (530b)$$

Introducing the approximation $\vec{\vec{\Phi}}_{\vec{A}\vec{B}}(\vec{r}t,\vec{r}'t') = \vec{\vec{\Phi}}_{\vec{A}\vec{B}}(\vec{r},\vec{r}';t-t')$, we find that Eq. (524) assumes the form

$$\langle \delta \hat{\vec{A}}(\vec{r},t)\rangle = \int_{t_0}^{t} dt' \int_{\mathcal{D}} dV' \vec{\vec{\Phi}}_{\vec{A}\vec{B}}(\vec{r},\vec{r}';t-t') \cdot \vec{F}_{\vec{B}}(\vec{r}',t') \tag{531}$$

given by linear response theory. In Kubo's formulation of linear response theory, the force $\vec{F}_{\vec{B}}(\vec{r},t)$ is assumed to be switched on in the infinite past $[t_0 = -\infty]$ when the system was in the state of thermal equilibrium. For this model, the above equation can be cast in the form of Eq. (528) for the case of a CW monochromatic force with $\vec{\vec{\chi}}_{\vec{A}\vec{B}}(\vec{k},\omega)$ denoting the generalized linear susceptibility

$$\vec{\vec{\chi}}_{\vec{A}\vec{B}}(\vec{k},\omega) = \lim_{\epsilon\to 0^+} \int_0^\infty dt \exp[+i(\omega+i\epsilon)t] \times \int_{\mathcal{D}} dV' \exp[+i\vec{k}\cdot(\vec{r}'-\vec{r})] \vec{\vec{\Phi}}_{\vec{A}\vec{B}}(\vec{r},\vec{r}';t). \tag{532}$$

For spatially homogeneous systems in the linear response regime, $\vec{\vec{\Phi}}_{\vec{A}\vec{B}}(\vec{r}\,t,\,\vec{r}'\,t') = \vec{\vec{\Phi}}_{\vec{A}\vec{B}}(\vec{r}-\vec{r}';t\,-\,t')$. Thus, the response at the space-time point $(\vec{r},t)$ to the disturbance at the space-time point $(\vec{r}',t')$ depends only on $\vec{r}-\vec{r}'$ and $t-t'$. A system possessing this character is said to be translationally invariant and temporally stationary.

Henceforth, we shall limit our considerations to spatially homogeneous systems in the linear response regime. For this case, the linear susceptibility assumes the simple form

$$\vec{\vec{\chi}}_{\vec{A}\vec{B}}(\vec{k},\omega) = \frac{i}{\hbar V} \lim_{\epsilon\to 0^+} \int_0^\infty dt \exp[+i(\omega+i\epsilon)t] \times \langle [\hat{\vec{A}}(\vec{k},t),\, \hat{\vec{B}}(-\vec{k},0)]_-\rangle_{\text{eq}}, \tag{533}$$

where

$$\hat{\vec{A}}(\vec{k},t) = \frac{1}{2}\sum_m [\hat{\vec{A}}_m(t),\, \exp(-i\vec{k}\cdot\hat{\vec{q}}_m(t))]_+ \tag{534}$$

and

$$\hat{\vec{B}}(-\vec{k},0) = \frac{1}{2}\sum_m [\hat{\vec{B}}_m(0),\, \exp(+i\vec{k}\cdot\hat{\vec{q}}_m(0))]_+. \tag{535}$$

One can demonstrate that causality implies that the real $\vec{\vec{\chi}}'_{\vec{A}\vec{B}}(\vec{k},\omega)$ and imaginary $\vec{\vec{\chi}}''_{\vec{A}\vec{B}}(\vec{k},\omega)$ parts of the linear susceptibility $\vec{\vec{\chi}}_{\vec{A}\vec{B}}(\vec{k},\omega)$ are connected by the Kramers–Kronig dispersion relations

$$\vec{\vec{\chi}}'_{\vec{A}\vec{B}}(\vec{k},\omega) = \frac{1}{\pi}\mathcal{P}\int_{-\infty}^{+\infty} d\omega' \frac{\vec{\vec{\chi}}''_{\vec{A}\vec{B}}(\vec{k},\omega')}{(\omega'-\omega)} \tag{536}$$

and

$$\vec{\vec{\chi}}''_{\vec{A}\vec{B}}(\vec{k},\omega) = -\frac{1}{\pi}\mathcal{P}\int_{-\infty}^{+\infty} d\omega' \frac{\vec{\vec{\chi}}'_{\vec{A}\vec{B}}(\vec{k},\omega')}{(\omega'-\omega)} \tag{537}$$

or

$$\vec{\vec{\chi}}'_{\vec{A}\vec{B}}(\vec{k},\omega) = \frac{2}{\pi}\mathcal{P}\int_0^\infty d\omega' \frac{\omega'\,\vec{\vec{\chi}}''_{\vec{A}\vec{B}}(\vec{k},\omega')}{(\omega'^2-\omega^2)} \tag{538}$$

and

$$\vec{\vec{\chi}}''_{\vec{A}\vec{B}}(\vec{k},\omega) = -\frac{2}{\pi}\mathcal{P}\int_0^\infty d\omega' \frac{\omega\,\vec{\vec{\chi}}'_{\vec{A}\vec{B}}(\vec{k},\omega')}{(\omega'^2-\omega^2)}. \tag{539}$$

In the above, the integrals are Cauchy principal values.

In the investigation of the response of a system subject to an external field, we are usually interested in the response $\langle \delta\vec{A}(\vec{r},t)\rangle$ of the property directly coupled to the field, i.e., $\langle \delta\vec{A}(\vec{r},t)\rangle = \langle \delta\vec{B}(\vec{r},t)\rangle$. For such investigations, the time-averaged power $Q(\omega)$ absorbed by the system from a CW monochromatic field over the period $2\pi/\omega$ is given by

$$Q(\omega) = \sum_{\vec{k}} Q(\vec{k},\omega), \tag{540}$$

where

$$Q(\vec{k},\omega) = \left(\frac{\omega V}{2}\right)\vec{F}_{\vec{B}}(\vec{k},\omega)\cdot\vec{\vec{\chi}}''_{\vec{B}\vec{B}}(\vec{k},\omega)\cdot\vec{F}^*_{\vec{B}}(\vec{k},\omega), \tag{541}$$

with

$$\vec{\vec{\chi}}''_{\vec{B}\vec{B}}(\vec{k},\omega) = \left(\frac{1}{\hbar V}\right)\tanh\left(\frac{\beta\hbar\omega}{2}\right)\int_{-\infty}^{+\infty} dt \exp(-i\omega t) \times \left\langle \frac{1}{2}[\hat{\vec{B}}(\vec{k},t),\, \hat{\vec{B}}(-\vec{k},0)]_+ \right\rangle_{\text{eq}}. \tag{542}$$

The above results connect the power dissipated by the system to spontaneous equilibrium fluctuations. In view of this connection, Eqs. (541) and (542) are said to represent a fluctuation–dissipation relation.

A simple problem for which the above-described formalism applies is the determination of the polarization $\langle \hat{\vec{p}}(\vec{r},t)\rangle$ of some dielectric medium subject to the electric field $\vec{E}(\vec{r},t)$. For this problem, $\hat{\vec{A}}(\vec{r}) = \hat{\vec{B}}(\vec{r}) = \hat{\vec{P}}(\vec{r})$ and $\vec{F}_{\vec{B}}(\vec{r},t) = \vec{E}(\vec{r},t)$, where the polarization operator $\hat{\vec{p}}(\vec{r})$ is defined by Eq. (515). Then $\vec{\vec{\chi}}_{\vec{A}\vec{B}}(\vec{k},\omega) = \vec{\vec{\chi}}_{\vec{P}\vec{P}}(\vec{k},\omega)$ represents the electric susceptibility of the system. If $\vec{\vec{\chi}}_{\vec{P}\vec{P}}(\vec{k},\omega)$ is determined using Eq. (529), $\vec{\vec{\chi}}_{\vec{P}\vec{P}}(\vec{k},\omega)$ represents a field-dependent nonlinear susceptibility. If Eq. (532) or (533) is used to determine $\vec{\vec{\chi}}_{\vec{P}\vec{P}}(\vec{k},\omega)$, this quantity represents a field-independent linear susceptibility.

Although we have focused our attention on quantum systems, the above formal development can be easily modified to handle classical systems. This can be accomplished by using the identity

$$\hat{L}_{\text{ext}}(t')|C\rangle = -\int_{\mathcal{D}} dV' |\{\vec{B}(\vec{r}',0),C\}\rangle \cdot \vec{F}_{\vec{B}}(\vec{r}',t')$$

for classical systems rather than the identity given by Eq. (522) for quantum systems, where $\{\vec{B}(\vec{r}',0),C\}$ is a Poisson bracket.

As one might anticipate from the difference in the results of the operation $\hat{L}_{\text{ext}}(t')|C\rangle$ for quantum and classical systems, passage to the classical limit can be

partially achieved by making replacements of the type $(i\hbar)|\,[A,\,B]_-\rangle \to |\{A,\,B\}\rangle$ or $(i\hbar)[\hat{A},\,\hat{B}]_- \to \{A,\,B\}$. Full recovery of the classical results from the various expressions also requires the following additional types of replacements: $\langle\hat{A}\rangle \to \langle A\rangle$, $\hat{\rho} \to \rho$, $\hat{\rho}_{\rm eq} \to \rho_{\rm eq}$, $\frac{1}{2}[\hat{A},\,\hat{B}]_+\rangle \to AB$, and $\hbar \to 0$.

C. Mori–Zwanzig Projection Operator Formalism

Since the characterization of a physical system usually requires only a limited number of average quantities or a contracted description of the system, Zwanzig suggested that we partition the density operator (probability density) appearing in the Liouville equation into relevant and irrelevant parts, with the relevant part bearing the relevant information for the physical system under consideration. This partitioning and the subsequent formal development are readily accomplished by working in the vector language introduced in Section 4, which allowed us to write the Liouville equation and the equation of motion for dynamical variables in the forms given by Eqs. (215) and (216), respectively. Here, we rewrite these equations as $(d/dt)|\rho_t\rangle = -\hat{L}\,|\rho_t\rangle$ and $(d/dt)\langle O_{j,t}| = -\langle O_{j,t}|\hat{L}$, where $\hat{L} = i\hat{\mathcal{L}}$ is called the transition operator.

The partitioning of the vector $|\rho_t\rangle$ corresponding to the density operator $\hat{\rho}(t)$ of some system into a relevant part, denoted by $|\rho_t^R\rangle$, and an irrelevant part, denoted by $|\rho_t^I\rangle$, can be accomplished by introducing the orthogonal projection operators $\hat{P}$ and $\hat{Q}$ possessing the properties $\hat{P}+\hat{Q}=\hat{I}$, $\hat{P}^2=\hat{P}$, $\hat{Q}^2=\hat{Q}$, and $\hat{P}\hat{Q}=\hat{Q}\hat{P}=\hat{0}$, where $\hat{0}$ is the null operator. Making use of the projection operators $\hat{P}$ and $\hat{Q}$, we write

$$|\rho_t\rangle = \big|\rho_t^R\big\rangle + \big|\rho_t^I\big\rangle, \tag{543}$$

where

$$\big|\rho_t^R\big\rangle = \hat{P}|\rho_t\rangle \tag{544}$$

and

$$\big|\rho_t^I\big\rangle = \hat{Q}|\rho_t\rangle. \tag{545}$$

Application of the projection operators $\hat{P}$ and $\hat{Q}$ to the Liouville equation $(d/dt)|\rho_t\rangle = -\hat{L}\,|\rho_t\rangle$ gives the following exact equation of motion for the relevant part $|\rho_t^R\rangle$ of $|\rho_t\rangle$:

$$\begin{aligned}\frac{d}{dt}\big|\rho_t^R\big\rangle &= -\hat{L}_{PP}\big|\rho_t^R\big\rangle \\ &\quad + \int_0^t dt'\,\hat{L}_{PQ}\exp[-\hat{L}_{QQ}(t-t')]\hat{L}_{QP}\big|\rho_{t'}^R\big\rangle \\ &\quad - \hat{L}_{PQ}\exp(-\hat{L}_{QQ}t)\,\hat{Q}\,|\rho_0\rangle,\end{aligned} \tag{546}$$

where $\hat{L}_{PP}$, $\hat{L}_{PQ}$, $\hat{L}_{QP}$, and $\hat{L}_{QQ}$ are the projected transition operators $\hat{L}_{PP}=\hat{P}\hat{L}\hat{P}$, $\hat{L}_{PQ}=\hat{P}\hat{L}\hat{Q}$, $\hat{L}_{QP}=\hat{Q}\hat{L}\hat{P}$, and $\hat{L}_{QQ}=\hat{Q}\hat{L}\hat{Q}$. The above equation of motion, first established by Zwanzig, is called the Zwanzig master equation.

According to Eq. (546), the vector $|\rho_t^R\rangle$ depends on itself at earlier times. Thus, $|\rho_t^R\rangle$ possesses memory. Equations of motion for quantities endowed with memory are labeled as non-Markovian.

The actual use of Eq. (546) requires us to specify the projection operator $\hat{P}$, which in turn depends on the problem at hand and how we wish to bias the final results. For example, suppose we want to construct a non-Markovian analogue of Eq. (492). The formal structure of Eqs. (492)–(494) suggests that the projection operator $\hat{P}$ be defined as

$$\hat{P} = \sum_{j,k} |\delta\tilde{O}_j\rho_{\rm eq}\rangle\chi_{jk}^{-1}\langle\delta O_k^\dagger\big|. \tag{547}$$

For the projection operator defined by Eq. (547), one can demonstrate that $\langle\delta O_j^\dagger|\rho_t\rangle = \langle\delta O_j^\dagger|\rho_t^R\rangle$. Then the formal expression $\langle\delta\hat{O}_j(t)\rangle = \langle\delta O_j^\dagger|\rho_t\rangle$ can be rewritten as $\langle\delta\hat{O}_j(t)\rangle = \langle\delta O_j^\dagger|\rho_t^R\rangle$. Making use of Eq. (546) to form the time derivative $(d/dt)\langle\delta O_j^\dagger|\rho_t^R\rangle$ in the relation $(d/dt)\langle\delta\hat{O}_j(t)\rangle = (d/dt)\langle\delta O_j^\dagger|\rho_t^R\rangle$ and subsequently introducing Eq. (547), we obtain the following non-Markovian analogue of Eq. (492):

$$\begin{aligned}\frac{d}{dt}\langle\delta\hat{O}_j(t)\rangle &= -\sum_k\Bigg[i\Omega_{jk}\langle\delta\hat{O}_k(t)\rangle \\ &\quad + \int_0^t dt'\mathbb{K}_{jk}(t-t')\langle\delta\hat{O}_k(t')\rangle\Bigg] + \mathbb{I}_j(t),\end{aligned} \tag{548}$$

where

$$i\Omega_{jk} = \sum_l \mathbb{M}_{jl}^S\chi_{lk}^{-1}, \tag{549}$$

$$\mathbb{K}_{jk}(t-t') = \sum_l \mathbb{M}_{jl}^C(t-t')\chi_{lk}^{-1}, \tag{550}$$

and

$$\mathbb{I}_j(t) = \big\langle\delta\dot{O}_j^\dagger\big|\,\hat{Q}\exp(-\hat{L}_{QQ}t)\hat{Q}\big|\rho_0\big\rangle, \tag{551}$$

with

$$\mathbb{M}_{jl}^S = \begin{cases} -\langle\delta O_l^\dagger|\delta\tilde{O}_j\rho_{\rm eq}\rangle \\ -\langle\delta\dot{O}_j^\dagger|\delta\tilde{O}_l\rho_{\rm eq}\rangle \\ \langle\delta O_j^\dagger|\delta\tilde{O}_l\rho_{\rm eq}\rangle \end{cases} \tag{552}$$

and

$$\mathbb{M}_{jl}^C(t-t') = \big\langle\delta\dot{O}_j^\dagger\big|\,\hat{Q}\exp[-\hat{L}_{QQ}(t-t')]\hat{Q}\big|\delta\tilde{O}_l\rho_{\rm eq}\big\rangle. \tag{553}$$

The first term in the sum on the right side of Eq. (548) describes the instantaneous response of the system to the displacements from thermal equilibrium. This response is characterized by the frequencies $\{\Omega_{jk}\}$. The second term in the sum on the right side of Eq. (548) describes memory effects. More specifically, it relates the displacements from equilibrium at time t to earlier values of these displacements through the memory kernels $\{\mathbb{K}_{jk}(t-t')\}$, which are commonly called memory functions. The last term on the right side of Eq. (548), denoted by $\mathbb{I}_j(t)$, is a source term describing effects due to the initial preparation of the system.

Application of the projection operators $\hat{P}$ and $\hat{Q}$ to the equation of motion $(d/dt)\langle\delta O_{j,t}^{\dagger}| = -\langle\delta O_{j,t}^{\dagger}|\hat{L}$ can be shown to give

$$\frac{d}{dt}\langle\delta O_{j,t}^{\dagger}| = -\sum_k \left[i\Omega_{jk}\langle\delta O_{k,t}^{\dagger}| + \int_0^t dt'\, \mathbb{K}_{jk}(t-t')\langle\delta O_{k,t'}^{\dagger}| \right] + \langle f_{j,t}|, \quad (554)$$

where

$$\langle f_{j,t}| = \langle\delta\dot{O}_j^{\dagger}| \exp(-\hat{L}_{QQ}t)\hat{Q}. \quad (555)$$

The above equation of motion, first obtained by Mori, is called a generalized Langevin equation due to its resemblance to the Langevin equations from classical Brownian motion theory. In essence, Eq. (554) describes the evolution of the dynamical vectors $\langle\delta O_{j,t}^{\dagger}|$ corresponding to the fluctuations $\delta\hat{O}_j(t) = \hat{O}_j(t) - \langle\hat{O}_j\rangle_{\text{eq}}$ about thermal equilibrium.

Making use of Eq. (554) to form the time derivative $(d/dt)\langle\delta O_{j,t}^{\dagger} \mid \rho_0\rangle$ in the relation $(d/dt)\langle\delta\hat{O}_j(t)\rangle = (d/dt)\langle\delta O_{j,t}^{\dagger} \mid \rho_0\rangle$, we again arrive at Eq. (548). Since Eq. (548) can be obtained from either Zwanzig's master equation or Mori's generalized Langevin equation, the projection operator approaches of Mori and Zwanzig are commonly referred to jointly as the Mori–Zwanzig projection operator formalism.

We should remark that Mori did not actually use the projection operator $\hat{P}$ defined by Eq. (547). Instead, Mori used the projection operator $\hat{P} = \sum_{jk} |\delta O_j\rangle \chi_{jk}^{-1} \langle\delta O_k|$. Also, Mori defined the inner product of $\langle\delta O_j|$ and $|\delta O_k\rangle$ by $\langle\delta O_j \mid \delta O_k\rangle = \text{Tr}\,\hat{\rho}_{\text{eq}}\delta\hat{O}_j\delta\hat{\bar{O}}_k$. The use of such projector operators and inner products tends to render the formalism less transparent. Moreover, the interpretation of the vectors $\langle\delta O_j|$ and $|\delta O_j\rangle$ and the mathematical relationship between them are unclear. This lack of clarity can lead to mathematical difficulties when subjecting the quantities $\mathbb{M}_{jk}^{C}(t)$ to detailed mathematical analysis.

In order to work with the set of non-Markovian equations of motion given by Eq. (548), we must first determine the quantities $\{\Omega_{jk}\}$, $\{\mathbb{K}_{jk}(t)\}$, and $\{\mathbb{I}_j(t)\}$ using some microscopic model as described by the transition operator $\hat{L}$ or the Liouville operator $\hat{\mathcal{L}}$. Given these quantities, we can proceed to solve the set of equations of motion for the displacements $\{\langle\delta\hat{O}_j(t)\rangle\}$ from thermal equilibrium. In general, this procedure for dealing with the non-Markovian equations of motion is difficult to implement.

In order to circumvent the difficulties in dealing with Eq. (548), investigators often introduce an ansatz for the memory functions, i.e., an analytic form is assumed for the memory functions on the basis of mathematical convenience and physical intuition. The simplest and most widely used memory function model is the Markovian approximation. In essence, one neglects non-Markovian retardation by assuming that the memory function $\mathbb{K}_{jk}(t)$ can be written as

$$\mathbb{K}_{jk}(t) = K_{jk}\delta(t), \quad (556)$$

where

$$K_{jk} = \sum_l M_{jl}^{C}\chi_{lk}^{-1}, \quad (557)$$

with

$$M_{jl}^{C} = \lim_{z\to 0^+} \langle\delta\dot{O}_j^{\dagger}|\hat{Q}(z\,\hat{Q} + \hat{L}_{QQ})^{-1}\hat{Q}|\delta\tilde{\hat{O}}_l\rho_{\text{eq}}\rangle. \quad (558)$$

In addition to making the Markovian approximation, the source term $\mathbb{I}_j(t)$ in Eq. (548) is usually neglected. This assumption is based on the idea that effects due to the initial preparation of the system are unimportant when the system is close to thermal equilibrium. A more rigorous justification for invoking this approximation can be obtained by using the maximum entropy principle to describe the initial state $|\rho_0\rangle$, which leads to the form $|\rho_0\rangle = |\exp[\Omega I - \beta\mathcal{H} - \sum_j \Lambda_j(0)O_j]\rangle$. Assuming that the system is initially close to thermal equilibrium, we can approximate $|\rho_0\rangle$ by $|\rho_{\text{eq}}\rangle + \sum_{j,k} |\delta\tilde{O}_j\rho_{\text{eq}}\rangle\chi_{jk}^{-1}\langle\delta\hat{O}_k(0)\rangle$. Making use of this approximate form for $|\rho_0\rangle$ to evaluate Eq. (551), we obtain $\mathbb{I}_j(t) = 0$.

Invoking the Markovian approximation and neglecting effects due to the initial preparation of the system, one finds that Eq. (548) assumes a form that is identical to the linear equation of motion given by Eq. (492) from QSM theory, except that the phenomenological time derivative is replaced by an instantaneous time derivative. The frequencies $\{\Omega_{jk}\}$ appearing in the Markovian approximation of Eq. (548) are identical to those appearing in Eq. (492). (The frequencies are also identical in the Markovian form.) Nonetheless, the formal expressions for K_{jk} differ due to the difference in the expressions for M_{jl}^{C}.

The classical limit of the quantum results can be obtained by making the replacements $\langle \delta O_j^{\dagger}| \to \langle \delta O_j^{*}|$ and $|\delta \tilde{O}_k \rho_{\rm eq}\rangle \to |\delta O_k \rho_{\rm eq}\rangle$. Apart from this modification, the classical and quantum results are identical.

As in the previously discussed theories of nonequilibrium processes, the time evolution of the displacements from equilibrium, as described by Eq. (548), is intimately connected to spontaneous equilibrium fluctuations. Such fluctuations are built into Eq. (548) through the projection operator $\hat{P}$ defined by Eq. (547).

Projection operators of the type given by Eq. (547) possess a global operational character in the sense that they operate on both the system of interest (system) and its surroundings (bath). An alternative approach to the problem of constructing contracted equations of motion is to use projection operators that operate only on the subspace spanned by the bath. Such bath projection operators have been used in conjunction with Zwanzig's master equation to construct equations of motion solely for the state vector $|\rho_t^S\rangle$ for the system of interest.

Apart from being used to construct contracted equations of motion, projection techniques have been used to develop powerful analytic and numerical tools that enable one to solve spectral and temporal problems without resorting to the solution of global equations of motion. In this respect, we mention Mori's memory function formalism and dual Lanczos transformation theory.

Mori's memory function formalism provides a framework for determining the time evolution of equilibrium autocorrelation functions (time correlation functions describing self-correlations) and their spectral transforms for the case of classical systems. The aforementioned ambiguity in the interpretation of the vectors $\langle \delta O_j|$ and $|\delta O_j\rangle$ in Mori's treatment of autocorrelation functions for quantum systems leads to mathematical difficulties.

Dual Lanczos transformation theory is a projection operator approach to nonequilibrium processes that was developed by the author to handle very general spectral and temporal problems. Unlike Mori's memory function formalism, dual Lanczos transformation theory does not impose symmetry restrictions on the Liouville operator and thus applies to both reversible and irreversible systems. Moreover, it can be used to determine the time evolution of equilibrium autocorrelation functions and cross-correlation functions (time correlation functions not describing self-correlations) and their spectral transforms for both classical and quantum systems. In addition, dual Lanczos transformation theory provides a number of tools for determining the temporal evolution of the averages of dynamical variables. Several years ago, it was demonstrated that the projection operator theories of Mori and Zwanzig represent special limiting cases of dual Lanczos transformation theory.

D. Maximum Entropy Approach to Nonequilibrium Processes

In Sections I.C and V.D it was shown that the basic results from equilibrium and nonequilibrium thermodynamics can be established from statistical mechanics by starting from the maximum entropy principle. The success of this approach to the formulation of equilibrium and nonequilibrium thermodynamics suggests that the maximum entropy principle can also be used to formulate a general theory of nonequilibrium processes that automatically includes the thermodynamic description of nonequilibrium systems. In this section, we formulate a theory possessing this character by making use of a time-dependent projection operator $\hat{P}(t)$ that projects the thermodynamic description $|\bar{\rho}_t\rangle$ of a system out of the global description $|\rho_t\rangle$ given by the solution of the Liouville equation. We shall refer to this theory as the maximum entropy approach to nonequilibrium processes.

The roots of the maximum entropy approach to nonequilibrium processes lie in Robertson's attempt to formulate a general method for constructing nonlinear equations of motion in nonequilibrium statistical mechanics, the Kawasaki–Gunton treatment of nonlinear transport processes in classical systems, and Grabert's treatment of nonequilibrium fluctuations and nonlinear transport processes in quantum systems. As with the maximum entropy approach, the Robertson, Kawasaki–Gunton, and Grabert theories of nonequilibrium processes employ a time-dependent projection operator that projects the thermodynamic description of a system out of the global description given by the solution of the classical or quantum Liouville equation. These theories also explicitly or implicitly assume the validity of the maximum entropy principle for systems out of equilibrium.

Although the maximum entropy approach and the Robertson, Kawasaki–Gunton, and Grabert theories employ the same basic methodology, they are not formally equivalent treatments of nonequilibrium processes. The time-dependent projection operator introduced by Robertson violates the conservation of probability. This defect was corrected for the case of classical systems with the time-dependent projection operator introduced by Kawasaki and Gunton. The time-dependent projection operator introduced by Grabert provides a suitable generalization of the Kawasaki–Gunton projection operator for quantum systems.

In the maximum entropy approach to nonequilibrium processes, we employ a time-dependent projection operator that is formally equivalent to the projection operator introduced by Grabert. However, it is written in such a way that it leads to results that are more transparent and more closely connected to the statistical mechanical

formulation of nonequilibrium thermodynamics, QSM theory, and Mori's projection operator formalism. This is accomplished by formulating the maximum entropy approach to nonequilibrium processes in the modern vector language introduced earlier in our discussion of the latter approaches to nonequilibrium processes. This vector language was not employed in the Grabert, Robertson, and Kawasaki–Gunton theories.

1. Global and Thermodynamic Descriptions of Nonequilibrium Systems

As in the Zwanzig approach to nonequilibrium processes, we partition the vector $|\rho_t\rangle$ corresponding to the density operator $\hat{\rho}(t)$ of some system into relevant and irrelevant parts by using orthogonal projection operators. However, the projection operators $\hat{P}(t)$ and $\hat{Q}(t)$ used to accomplish this task are time dependent rather than time independent. Nonetheless, $\hat{P}(t)$ and $\hat{Q}(t)$ possess the usual properties $\hat{P}(t)+\hat{Q}(t)=\hat{I}$, $\hat{P}(t)^2=\hat{P}(t)$, $\hat{Q}(t)^2=\hat{Q}(t)$, and $\hat{P}(t)\hat{Q}(t)=\hat{Q}(t)\hat{P}(t)=\hat{0}$ of orthogonal projection operators, where $\hat{0}$ is the null operator.

We require the projection operator $\hat{P}(t)$ to be such that it projects the thermodynamic description $|\bar{\rho}_t\rangle$ of a system out of the global description $|\rho_t\rangle$, i.e., $|\bar{\rho}_t\rangle=\hat{P}(t)|\rho_t\rangle$. The global description $|\rho_t\rangle$ of the system is given by the solution of the Liouville equation $(d/dt)\,|\rho_t\rangle=-\hat{L}\,|\rho_t\rangle$, where $\hat{L}=i\hat{\mathcal{L}}$. The thermodynamic description $|\bar{\rho}_t\rangle$ of the system is obtained by employing the time-dependent version of the maximum entropy principle.

As in our discussions of nonequilibrium thermodynamics and QSM theory, the thermodynamic description of a system is defined by specifying the values of certain thermodynamic coordinates. The thermodynamic coordinates should include the average energy $\langle\hat{\mathcal{H}}\rangle$ or the local energy density $\langle\hat{\rho}_{\mathcal{H}}(\vec{r},t)\rangle$ and time-dependent external parameters, such as a variable volume $\langle V(t)\rangle$ or electric field $\langle\hat{\vec{E}}(\vec{r},t)\rangle$. Independent of the nature of the thermodynamic coordinates, they must be linearly independent in order for $|\bar{\rho}_t\rangle$ to be unique.

We adopt the point of view that the system is actually a composite system made up of the system of interest and its surroundings. As indicated in our discussion of QSM theory, the system of interest and its surroundings do not necessarily occupy different spatial domains.

The global and thermodynamic descriptions of a system are required to be equivalent at the time origin and to give the same results for the values of the thermodynamic coordinates at any given instant in time. Also, we require the global and thermodynamic descriptions to satisfy the conservation of probability. These requirements can be expressed as follows: $|\rho_{t_0}\rangle=|\bar{\rho}_{t_0}\rangle$, where t_0 is the time origin; $\langle\hat{O}_j(t)\rangle=\langle O_j^\dagger|\rho_t\rangle=\langle O_j^\dagger|\bar{\rho}_t\rangle$ for spatially independent thermodynamic coordinates and $\langle\hat{O}_j(\vec{r},t)\rangle=\langle O_j^\dagger(\vec{r})|\rho_t\rangle=\langle O_j^\dagger(\vec{r})|\bar{\rho}_t\rangle$ for spatially dependent thermodynamic coordinates; and $\langle I|\rho_t\rangle=\langle I|\bar{\rho}_t\rangle=1$ for all t. For the sake of simplicity, we take $t_0=0$.

As in our earlier discussions of nonequilibrium systems, we restrict our considerations to systems that are characterized by either spatially independent thermodynamic coordinates $\{\langle\hat{O}_j(t)\rangle\}$ or spatially dependent thermodynamics coordinates $\{\langle\hat{O}_j(\vec{r},t)\rangle\}$. The generalization of the results to include both types of thermodynamic coordinates can be accomplished in a straightforward fashion.

2. Spatially Independent Thermodynamic Coordinates

For the case of spatially independent thermodynamic coordinates $\{\langle\hat{O}_j(t)\rangle\}$, we project the thermodynamic description $|\bar{\rho}_t\rangle$ out of the global description $|\rho_t\rangle$ with the projection operator

$$\hat{P}(t)=|\bar{\rho}_t\rangle\langle I|+\sum_{j,k}|\delta_t\tilde{O}_j\bar{\rho}_t\rangle\chi_{jk}^{-1}(t)\langle\delta_t O_k^\dagger|. \tag{559}$$

In Eq. (559), $\chi_{jk}^{-1}(t)$ denotes a matrix element of the inverse $\boldsymbol{\chi}^{-1}(t)$ of the nonequilibrium susceptibility matrix $\boldsymbol{\chi}(t)$, with the matrix elements $\chi_{jk}(t)$ of $\boldsymbol{\chi}(t)$ defined by Eqs. (360a)–(363). In the vector language adopted here, we write

$$\chi_{jk}(t)=\langle\delta_t O_k^\dagger|\delta_t\tilde{O}_j\bar{\rho}_t\rangle. \tag{560}$$

The vectors $|\delta_t\tilde{O}_j\bar{\rho}_t\rangle$ and $\langle\delta_t O_k^\dagger|$ appearing in Eqs. (559) and (560) correspond to the operators $\delta_t\hat{\tilde{O}}_j\hat{\bar{\rho}}(t)$ and $\delta_t\hat{O}_k$, where $\hat{\bar{\rho}}(t)$ is the statistical density operator given by Eq. (341), and $\delta_t\hat{\tilde{O}}_j$ and $\delta_t\hat{O}_k$ are the operators defined by Eqs. (361) and (362), respectively. Unless stated otherwise, the tilde in the various expressions discussed here and in the remainder of our discussion of the maximum entropy approach to nonequilibrium processes should be interpreted as a generalized Kubo transform as defined by Eq. (363) or (450). As evident form Eq. (363), the generalized Kubo transform carries an additional time dependence. For the sake of notational simplicity, we have suppressed it.

If the system is close to thermal equilibrium, the projection operator $\hat{P}(t)$ assumes the time-independent form

$$\hat{P}=|\rho_{\rm eq}\rangle\langle I|+\sum_{j,k}|\delta\tilde{O}_j\rho_{\rm eq}\rangle\chi_{jk}^{-1}\langle\delta O_k^\dagger|, \tag{561}$$

where the tilde indicates the usual time-independent Kubo transform $|\delta\tilde{O}_j\rho_{\rm eq}\rangle=\int_0^1 d\lambda|\rho_{\rm eq}^\lambda\delta O_j\rho_{\rm eq}^{-\lambda}\rho_{\rm eq}\rangle$ and χ_{jk}^{-1} is a matrix element of the inverse $\boldsymbol{\chi}^{-1}$ of the equilibrium susceptibility matrix $\boldsymbol{\chi}$, with the matrix elements χ_{jk} of $\boldsymbol{\chi}$ given by $\chi_{jk}=\langle\delta O_k^\dagger\,|\,\delta\tilde{O}_j\rho_{\rm eq}\rangle$. The second term on the

right side of the above equation corresponds to the time-independent projection operator $\hat{P}$ adopted in our discussion of the Mori–Zwanzig projection operator formalism. [See Eq. (547).]

One can show that the application of the projection operators $\hat{P}(t)$ and $\hat{Q}(t)$ to the Liouville equation $(d/dt)|\rho_t\rangle = -\hat{L}|\rho_t\rangle$ gives the following exact equation of motion for the thermodynamic description $|\bar{\rho}_t\rangle$ of the system:

$$\frac{d}{dt}|\bar{\rho}_t\rangle = -\hat{L}_{PP}(t)|\bar{\rho}_t\rangle + \int_0^t dt' \hat{L}_{PQ}(t5)\hat{\Phi}_{QQ}(t,t')\hat{L}_{QP}(t')|\bar{\rho}_{t'}\rangle. \quad (562)$$

The quantities $\hat{L}_{PP}(t)$, $\hat{L}_{PQ}(t)$, and $\hat{L}_{QP}(t)$ are the projbreak ected transition operators $\hat{L}_{PP}(t) = \hat{P}(t)\hat{L}\hat{P}(t)$, $\hat{L}_{PQ}(t) = \hat{P}(t)\hat{L}\hat{Q}(t)$, and $\hat{L}_{QP}(t) = \hat{Q}(t)\hat{L}\hat{P}(t)$. It also convenient to define $\hat{L}_{QQ}(t) = \hat{Q}(t)\hat{L}\hat{Q}(t)$.

In Eq. (562), the quantity $\hat{\Phi}_{QQ}(t,t')$ is a projected propagator given by the solution of the equation of motion

$$\frac{\partial}{\partial t}\hat{\Phi}_{QQ}(t,t') = -\hat{L}_{QQ}(t)\hat{\Phi}_{QQ}(t,t') \quad (563)$$

subject to the boundary condition $\hat{\Phi}_{QQ}(t',t') = \hat{Q}(t')$. The formal solution of this equation can be expressed in terms of a positive time-ordered exponential:

$$\hat{\Phi}_{QQ}(t,t') = \exp_+\left[-\int_{t'}^t d\tau \hat{L}_{QQ}(\tau)\right]\hat{Q}(t'). \quad (564)$$

From the above, we see that the projected propagator $\hat{\Phi}_{QQ}(t,t')$ is an operator-valued functional of the mean history $\{\langle\hat{O}_j(\tau)\rangle; t' \le \tau \le t\}$ of the system.

An equation of motion identical in form to Eq. (562) was first obtained by Robertson. Nonetheless, Robertson's equation of motion is faulty due to the use of the defective projection operator $\hat{P}(t) = \sum_j [\partial|\bar{\rho}_t\rangle/\partial\langle\hat{O}_j(t)\rangle]\langle O_j^\dagger|$, which violates the conservation of probability requirement $\langle I|\rho_t\rangle = \langle I|\bar{\rho}_t\rangle = 1$.

The faulty equation of motion obtained by Robertson was first corrected by Kawasaki and Gunton in a treatment of transport processes in classical systems. These investigators used a projection operator that corresponds to the classical limit of Eq. (559). The classical equation of motion obtained by Kawasaki and Gunton and the more general result given by Eq. (562) do not violate the conservation of probability requirement.

It should be noted that the classical limit of Eq. (559) can be obtained by making the replacements $\langle\delta_t O_k^\dagger| \to \langle\delta_t O_k^*|$ and $|\delta_t\tilde{O}_j\bar{\rho}_t\rangle \to |\delta_t O_j\bar{\rho}_t\rangle$ in Eqs. (559) and (560). This rule for passing to the classical limit can be applied to all of the expressionsdiscussed above and throughout the remainder of our discussion of the maximum entropy approach to nonequilibrium processes.

The evolution of thermodynamic description $|\bar{\rho}_t\rangle$ of the system is connected to the evolution of its global description $|\rho_t\rangle$ through the relation

$$|\rho_t\rangle = |\bar{\rho}_t\rangle - \int_0^t dt' \hat{\Phi}_{QQ}(t,t')\hat{L}_{QP}(t')|\bar{\rho}_{t'}\rangle. \quad (565)$$

This result was obtained by Grabert in a treatment of nonequilibrium fluctuations and transport processes in quantum systems. In order to obtain this result, Grabert employed the projection operator

$$\hat{P}(t) = |\bar{\rho}_t\rangle\langle I| + \sum_j [\partial|\bar{\rho}_t\rangle/\partial\langle\hat{O}_j(t)\rangle] \times \left[\langle O_j^\dagger| - \langle\hat{O}_j(t)\rangle\langle I|\right],$$

which can be shown to be equivalent to the projection operator $\hat{P}(t)$ defined by Eq. (559).

The generalized master equation given by Eq. (562) can be used to construct the following generalization of the phenomenological equations from nonequilibrium thermodynamics for the case of spatially independent thermodynamic coordinates:

$$\frac{d}{dt}\langle\hat{O}_j(t)\rangle = \langle\dot{\hat{O}}_j\rangle_{\bar{\rho}(t)} + \sum_k \int_0^t dt' M_{jk}^C(t,t')\Lambda_k(t') \quad (566a)$$

$$= \sum_k \left[M_{jk}^S(t)\Lambda_k(t) + \int_0^t dt' M_{jk}^C(t,t')\Lambda_k(t')\right], \quad (566b)$$

where

$$\langle\dot{\hat{O}}_j\rangle_{\bar{\rho}(t)} = \langle\dot{O}_j^\dagger|\bar{\rho}_t\rangle, \quad (567)$$

$$M_{jk}^S(t) = \langle O_j^\dagger|\tilde{O}_k\bar{\rho}_t\rangle, \quad (568)$$

and

$$M_{jk}^C(t,t') = \langle\dot{O}_j^\dagger|\hat{Q}(t)\hat{\Phi}_{QQ}(t,t')\hat{Q}(t')|\tilde{O}_k\bar{\rho}_{t'}\rangle. \quad (569)$$

The generalized Kubo transform in $|\tilde{O}_k\bar{\rho}_t\rangle[|\tilde{O}_k\bar{\rho}_{t'}\rangle]$ depends on t [t'].

Unlike the phenomenological equations from nonequilibrium thermodynamics, Eqs. (566a) and (566b) are non-Markovian. Moreover, $(d/dt)\langle\hat{O}_j(t)\rangle$ has a nonlinear dependence on the thermodynamic parameters $\{\Lambda_j(t)\}$. This nonlinear dependence is carried by the nonequilibrium time correlation functions $M_{jk}^S(t)$ and $M_{jk}^C(t,t')$, defined by Eqs. (568) and (569), respectively. Since the thermodynamic parameters $\{\Lambda_j(t)\}$ depend on the thermodynamic

coordinates $\{\langle \hat{O}_j(t)\rangle\}$, Eqs. (566a) and (566b) can be regarded as nonlinear, non-Markovian equations of motion for the thermodynamic coordinates.

Generalized phenomenological equations of the form of Eq. (566a) were first obtained by Robertson using the aforementioned defective projection operator. Suitably corrected generalized phenomenological equations were first obtained by Kawasaki and Gunton for classical systems and by Grabert for quantum systems. Nonetheless, these investigators neither adopted the modern vector language used in Eqs. (567)–(569) nor established Eq. (566b). These results were established by the author.

If the system is close to thermal equilibrium, we can approximate Eq. (566b) by

$$\frac{d}{dt}\langle \hat{O}_j(t)\rangle = \sum_k \left[M^S_{jk}\Lambda_k(t) + \int_0^t dt' M^C_{jk}(t-t')\Lambda_k(t') \right], \quad (570)$$

where

$$M^S_{jk} = \begin{cases} -\langle O^\dagger_k | \tilde{\hat{O}}_j \rho_{\rm eq}\rangle \\ -\langle \dot{O}^\dagger_j | \tilde{O}_k \rho_{\rm eq}\rangle \\ \langle O^\dagger_j | \tilde{\hat{O}}_k \rho_{\rm eq}\rangle \end{cases} \quad (571)$$

and

$$M^C_{jk}(t-t') = \langle \dot{O}^\dagger_j | \hat{Q} \exp[-\hat{L}_{QQ}(t-t')]\hat{Q} | \tilde{\hat{O}}_k \rho_{\rm eq}\rangle, \quad (572)$$

with $\hat{L}_{QQ}$ denoting the time-independent projected transition operator $\hat{L}_{QQ} = \hat{Q}\hat{L}\hat{Q}$. The projection operator $\hat{Q}$ is the complement of the time-independent projection operator $\hat{P}$ defined by Eq. (561). [Since $\hat{L}|\rho_{\rm eq}\rangle = |0\rangle$ and $\langle I|\hat{L} = \langle 0|$, where $\langle 0|$ and $|0\rangle$ are null vectors, $\hat{Q}$ can be replaced by the complement of the projection operator $\hat{P}$ defined by Eq. (547).] In Eqs. (571) and (572), the tilde indicates the usual Kubo transform for a system in thermal equilibrium.

From the above, we see that the nonlinear, non-Markovian generalized phenomenological equations (566a) and (566b) become linear, non-Markovian equations (570), in the regime close to thermal equilibrium. In the Markovian approximation, the linear non-Markovian equations assume a form that is identical to the linear phenomenological equations given by (484) from the linear domain of QSM theory, except the phenomenological time derivative is replaced by an instantaneous time derivative.

As for Eq. (484), the phenomenological coefficient M_{jk} resulting from the Markovian approximation of Eq. (570) is given by Eq. (486), with M^S_{jk} given by Eq. (488) or (497). However, M^C_{jk} is given by

$$M^C_{jk} = \lim_{z\to 0^+} \langle \delta\dot{O}^\dagger_j | \hat{Q}(z\,\hat{Q} + \hat{L}_{QQ})^{-1}\hat{Q} | \delta\tilde{\hat{O}}_k \rho_{\rm eq}\rangle \quad (573)$$

rather than Eq. (489), (495), (498), or (499).

In writing Eq. (573), we made use of the identities $\langle \delta\dot{O}^\dagger_j| = \langle \dot{O}^\dagger_j|$ and $|\delta\tilde{\hat{O}}_k\rho_{\rm eq}\rangle = |\tilde{\hat{O}}_k\rho_{\rm eq}\rangle$. It should be noted that the expressions given by (497) and (571) are equivalent.

Fluctuations $\delta\hat{O}_j(t) = \hat{O}_j(t) - \langle \hat{O}_j(t)\rangle$ about the mean path, defined by $\langle \hat{O}_j(t)\rangle$, can be described by the vectors $\langle \delta O^\dagger_{j,t}| = \langle O^\dagger_{j,t}| - \langle \hat{O}_j(t)\rangle\langle I|$. The time derivative of these vectors is given by $(d/dt)\langle \delta O^\dagger_{j,t}| = \langle \dot{O}^\dagger_{j,t}| - \langle \dot{\hat{O}}_j(t)\rangle\langle I|$. Application of the projection operators $\hat{P}(t)$ and $\hat{Q}(t)$ to the latter equation can be shown to give

$$\frac{d}{dt}\langle \delta O^\dagger_{j,t}| = -\sum_k \left[i\Omega_{jk}(t)\langle \delta O^\dagger_{k,t}| + \int_0^t dt' \mathbb{K}_{jk}(t,t')\langle \delta O^\dagger_{k,t'}| \right] + \langle f_{j,t}|, \quad (574)$$

where

$$i\Omega_{jk}(t) = \sum_l \mathbb{M}^S_{jl}(t)\chi^{-1}_{lk}(t), \quad (575)$$

$$\mathbb{K}_{jk}(t,t') = \sum_l \mathbb{M}^C_{jl}(t,t')\chi^{-1}_{lk}(t'), \quad (576)$$

and

$$\langle f_{j,t}| = -\langle \delta_t O^\dagger_j | \hat{L}\hat{Q}(t)\hat{\Phi}_{QQ}(t,0) - \int_0^t dt' \langle \delta_t O^\dagger_j | \hat{L}\hat{Q}(t)\hat{\Phi}_{QQ}(t,t')\hat{Q}(t')\exp(-\hat{L}t'), \quad (577)$$

with

$$\mathbb{M}^S_{jl}(t) = \langle \delta_t O^\dagger_j | \hat{L} | \delta_t \tilde{O}_l \bar{\rho}_t\rangle \quad (578)$$

and

$$\mathbb{M}^C_{jl}(t,t') = -\langle \delta_t O^\dagger_j | \hat{L}\hat{Q}(t)\hat{\Phi}_{QQ}(t,t')\hat{Q}(t')\hat{L} | \delta_{t'}\tilde{O}_l\bar{\rho}_{t'}\rangle. \quad (579)$$

The generalized Kubo transform in $|\delta_t\tilde{O}_l\bar{\rho}_t\rangle$ $[|\delta_{t'}\tilde{O}_l\bar{\rho}_{t'}\rangle]$ depends on t $[t']$.

The equation of motion given by Eq. (574) is an exact generalized Langevin equation for the vectors $\langle \delta O^\dagger_{j,t}|$ corresponding to the fluctuations $\delta\hat{O}_j(t) = \hat{O}_j(t) - \langle \hat{O}_j(t)\rangle$ about the mean path. An equation of motion of this form was first established by Grabert in the aforementioned treatment of nonequilibrium fluctuations and transport processes in quantum systems. Nonetheless, Grabert neither employed the modern vector language adopted here

nor established the transparent formal expressions given by Eqs. (575)–(579). These results were established by the author.

The generalized Langevin equation given by Eq. (554) in our discussion of the Mori–Zwanzig projection operator formalism is an equation of motion for the vectors $\langle \delta O_{j,t}^{\dagger}|$ corresponding to the fluctuations $\delta \hat{O}_j(t) = \hat{O}_j(t) - \langle \hat{O}_j \rangle_{\rm eq}$ about thermal equilibrium. If the vector $\langle \delta O_{j,t}^{\dagger}|$ in the generalized Langevin equation given by Eq. (574) is given the same interpretation, it reduces to the generalized Langevin equation given by Eq. (554).

With Eq. (574) at our disposal, we can consider the time evolution of the displacements $\langle \delta \hat{O}_j(t) \rangle = \langle \hat{O}_j(t) \rangle - \langle \hat{O}_j \rangle_{\rm eq}$ from thermal equilibrium. Making use of Eq. (574) to form the time derivative $(d/dt)\langle \delta O_{j,t}^{\dagger}|\rho_{\rm eq}\rangle$ in the expression $(d/dt)\langle \delta \hat{O}_j(t)\rangle = -(d/dt)\langle \delta O_{j,t}^{\dagger}|\rho_{\rm eq}\rangle$, we obtain

$$\frac{d}{dt}\langle \delta \hat{O}_j(t)\rangle = -\sum_k \left[i\Omega_{jk}(t)\langle \delta \hat{O}_k(t)\rangle + \int_0^t dt'\, \mathbb{K}_{jk}(t,t')\langle \delta \hat{O}_k(t')\rangle \right] + \mathbb{I}_j(t), \quad (580)$$

where

$$\mathbb{I}_j(t) = \left\langle \delta_t O_j^{\dagger} \middle| \hat{L}\,\hat{Q}(t)\hat{\Phi}_{QQ}(t,0) \middle| \rho_{\rm eq} \right\rangle + \int_0^t dt' \left\langle \delta_t O_j^{\dagger} \middle| \hat{L}\,\hat{Q}(t)\hat{\Phi}_{QQ}(t,t')\hat{Q}(t') \middle| \rho_{\rm eq} \right\rangle. \quad (581)$$

If the system is close to thermal equilibrium, Eq. (580) assumes the form of the non-Markovian equation of motion given by Eq. (548) in our discusssion of the Mori–Zwanzig projection operator formalism.

In order to work with the various equations of motion discussed above, one must be able to determine the thermodynamic parameters $\{\Lambda_j(t)\}$. These quantities can be determined by solving the set of equations (346), which is equivalent to solving the basic equations $\langle \hat{O}_j(t)\rangle = {\rm Tr}\,\hat{\rho}(t)\hat{O}_j$ defining the thermodynamic state of the system, where $\hat{\rho}(t)$ is given by Eq. (341).

Substituting Eq. (566b) into Eq. (365), we obtain the following closed set of nonlinear equations of motion for the thermodynamic parameters:

$$\frac{d}{dt}\Lambda_j(t) = -\sum_k \left[\mathfrak{N}_{jk}^{S}(t)\Lambda_k(t) + \int_0^t dt'\, \mathfrak{N}_{jk}^{C}(t,t')\Lambda_k(t') \right], \quad (582)$$

where

$$\mathfrak{N}_{jk}^{S}(t) = \sum_l \chi_{jl}^{-1}(t) M_{lk}^{S}(t) \quad (583)$$

and

$$\mathfrak{N}_{jk}^{C}(t,t') = \sum_l \chi_{jl}^{-1}(t) M_{lk}^{C}(t,t'). \quad (584)$$

Given the solution of the set of equations of motion given by Eq. (582), we can determine the time evolution of the thermodynamic parameters $\{\Lambda_j(t)\}$. This not only enables us to determine the quantities appearing in the various equations of motion discussed above, but also enables us to determine the time evolution of the thermodynamic coordinates $\{\langle \hat{O}_j(t)\rangle\}$. The latter task can be accomplished by using the formal expression $\langle \hat{O}_j(t)\rangle = {\rm Tr}\,\hat{\rho}(t)\hat{O}_j$, where $\hat{\rho}(t)$ is given by Eq. (341). Alternatively, one can actually solve the relevant equations of motion for the thermodynamic coordinates.

3. Spatially Dependent Thermodynamic Coordinates

For the case of spatially dependent thermodynamic coordinates $\{\langle \hat{O}_j(\vec{r},t)\rangle\}$, we project the thermodynamic description $|\bar{\rho}_t\rangle$ out of the global description $|\rho_t\rangle$ with the projection operator

$$\hat{P}(t) = |\bar{\rho}_t\rangle\langle I| + \sum_{j,k}\int_{\mathcal{D}} dV \int_{\mathcal{D}} dV' |\delta_t \tilde{O}_j(\vec{r})\bar{\rho}_t\rangle \times \chi_{jk}^{-1}(\vec{r},\vec{r}';t)\langle \delta_t O_k^{\dagger}(\vec{r}')|. \quad (585)$$

In Eq. (585), $\chi_{jk}^{-1}(\vec{r},\vec{r}';t)$ denotes a matrix element of the inverse $\boldsymbol{\chi}^{-1}(t)$ of the nonequilibrium susceptibility matrix $\boldsymbol{\chi}(t)$, with the matrix elements $\chi_{jk}(\vec{r},\vec{r}';t)$ of $\boldsymbol{\chi}(t)$ defined by Eqs. (447a)–(450). In the vector language adopted here,

$$\chi_{jk}(\vec{r},\vec{r}';t) = \langle \delta_t O_k^{\dagger}(\vec{r}')|\delta_t \tilde{O}_j(\vec{r})\bar{\rho}_t\rangle. \quad (586)$$

The vectors $|\delta_t \tilde{O}_j(\vec{r})\bar{\rho}_t\rangle$ and $\langle \delta_t O_k^{\dagger}(\vec{r}')|$ appearing in Eqs. (585) and (586) correspond to the operators $\delta_t \hat{\tilde{O}}_j(\vec{r}\,)\hat{\rho}(t)$ and $\delta_t \hat{O}_k(\vec{r}')$, where $\hat{\rho}(t)$ is the statistical density operator given by Eq. (425), and $\delta_t \hat{\tilde{O}}_j(\vec{r}\,)$ and $\delta_t \hat{O}_k(\vec{r}')$ are the operators defined by Eqs. (448) and (449), respectively. As for the case of spatially independent thermodyamic coordinates, the tilde is used to indicate a generalized Kubo transform, which carries an additional time dependence. [See Eq. (450).]

If the system is close to thermal equilibrium, the projection operator $\hat{P}(t)$ assumes the time-independent form

$$\hat{P} = |\rho_{\rm eq}\rangle\langle I| + \sum_{j,k}\int_{\mathcal{D}} dV \int_{\mathcal{D}} dV' |\delta\tilde{O}_j(\vec{r})\rho_{\rm eq}\rangle \times \chi^{-1}_{jk}(\vec{r},\vec{r}')\langle\delta O^{\dagger}_k(\vec{r}')|, \quad (587)$$

where the tilde indicates the usual time-independent Kubo transform $|\delta\tilde{O}_j(\vec{r})\rho_{\rm eq}\rangle = \int_0^1 d\lambda\, |\rho_{\rm eq}^{\lambda}\delta O_j(\vec{r})\rho_{\rm eq}^{-\lambda}\rho_{\rm eq}\rangle$ and $\chi^{-1}_{jk}(\vec{r},\vec{r}')$ is a matrix element of the inverse $\boldsymbol{\chi}^{-1}$ of the equilibrium susceptibility matrix $\boldsymbol{\chi}$, with the matrix elements $\chi_{jk}(\vec{r},\vec{r}')$ of $\boldsymbol{\chi}$ given by $\chi_{jk}(\vec{r},\vec{r}') = \langle\delta O^{\dagger}_k(\vec{r}')|\delta\tilde{O}_j(\vec{r})\rho_{\rm eq}\rangle$.

Making use of the projection operator defined by Eq. (585) and the generalized master equation given by Eq. (562), one can construct the following generalized phenomenological equations for the case of spatially dependent thermodynamic coordinates:

$$\frac{\partial}{\partial t}\langle\hat{O}_j(\vec{r},t)\rangle = \langle\hat{\hat{O}}_j(\vec{r})\rangle_{\bar{\rho}(t)} + \sum_k\int_{\mathcal{D}} dV' \times \int_0^t dt' M^C_{jk}(\vec{r}t,\vec{r}'t')\Lambda_k(\vec{r}',t') \quad (588a)$$

$$= \sum_k\left[\int_{\mathcal{D}} dV' M^S_{jk}(\vec{r},\vec{r}';t)\Lambda_k(\vec{r}',t) + \int_{\mathcal{D}} dV'\int_0^t dt' M^C_{jk}(\vec{r}t,\vec{r}'t')\Lambda_k(\vec{r}',t')\right], \quad (588b)$$

where

$$\langle\hat{\hat{O}}_j(\vec{r})\rangle_{\bar{\rho}(t)} = \left\langle\dot{O}^{\dagger}_j(\vec{r})\right|\bar{\rho}_t\rangle, \quad (589)$$

$$M^S_{jk}(\vec{r},\vec{r}';t) = \left\langle O^{\dagger}_j(\vec{r})\right|\tilde{\tilde{O}}_k(\vec{r}')\bar{\rho}_t\rangle, \quad (590)$$

and

$$M^C_{jk}(\vec{r}\,t,\vec{r}'\,t') = \left\langle\dot{O}^{\dagger}_j(\vec{r})\right|\hat{Q}(t)\hat{\Phi}_{QQ}(t,t')\hat{Q}(t')\left|\tilde{\tilde{O}}_k(\vec{r}')\bar{\rho}_{t'}\right\rangle. \quad (591)$$

The generalized Kubo transform in $|\tilde{\tilde{O}}_k(\vec{r}')\bar{\rho}_t\rangle$ $[\tilde{\tilde{O}}_k(\vec{r}')\bar{\rho}_{t'}\rangle]$ depends on t $[t']$.

The exact phenomenological equations given by Eqs. (588a) and (588b) represent spatially dependent generalizations of the phenomenological equations given by Eqs. (566a) and (566b). Similar to the case of spatially independent thermodynamic coordinates, we find that the rate of change $(\partial/\partial t)\langle\hat{O}_j(\vec{r},t)\rangle$ of the thermodynamic coordinate $\langle\hat{O}_j(\vec{r},t)\rangle$ has a nonlinear dependence on the thermodynamic parameters $\{\Lambda_j(\vec{r},\,t)\}$ throughout the spatial domain $\mathcal{D}$. This nonlinear dependence is carried by the nonequilibrium time correlation functions $M^S_{jk}(\vec{r},\vec{r}';t)$ and $M^C_{jk}(\vec{r}\,t,\vec{r}'\,t')$, defined by Eqs. (590) and (591), respectively. Since the thermodynamic parameters $\{\Lambda_j(\vec{r},\,t)\}$ depend on the thermodynamic coordinates $\{\langle\hat{O}_j(\vec{r},t)\rangle\}$ throughout the spatial domain $\mathcal{D}$, Eqs. (588a) and (588b) can be regarded as nonlinear, non-Markovian integral equations of motion for the thermodynamic coordinates.

Generalized phenomenological equations of the form of Eq. (588a) were first obtained by Robertson using the defective projection operator $\hat{P}(t) = \sum_j[\delta|\bar{\rho}_t\rangle/\delta\langle\hat{O}_j(\vec{r},t)\rangle]\langle O^{\dagger}_j(\vec{r},t)|$, which violates the conservation of probability requirement $\langle I|\rho_t\rangle = \langle I|\bar{\rho}_t\rangle = 1$. The results given by Eqs. (588a) and (588b) do not suffer from this difficulty.

If the system is close to thermal equilibrium, we can approximate Eq. (588b) by

$$\frac{\partial}{\partial t}\langle\hat{O}_j(\vec{r},t)\rangle = \sum_k\left[\int_{\mathcal{D}} dV' M^S_{jk}(\vec{r},\vec{r}')\Lambda_k(\vec{r}',t) + \int_{\mathcal{D}} dV'\int_0^t dt' M^C_{jk}(\vec{r},\vec{r}';t-t')\Lambda_k(\vec{r}',t')\right], \quad (592)$$

where

$$M^S_{jk}(\vec{r},\vec{r}\,') = \begin{cases} -\left\langle O^{\dagger}_k(\vec{r}')\right|\tilde{\tilde{O}}_j(\vec{r})\rho_{\rm eq}\rangle \\ -\left\langle\dot{O}^{\dagger}_j(\vec{r})\right|\tilde{O}_k(\vec{r}')\rho_{\rm eq}\rangle \\ \left\langle O^{\dagger}_j(\vec{r})\right|\tilde{\tilde{O}}_k(\vec{r}')\rho_{\rm eq}\rangle \end{cases} \quad (593)$$

and

$$M^C_{jk}(\vec{r},\vec{r}';t-t') = \left\langle\dot{O}^{\dagger}_j(\vec{r})\right|\hat{Q}\exp[-\hat{L}_{QQ}(t-t')] \times \hat{Q}|\tilde{\tilde{O}}_k(\vec{r}')\rho_{\rm eq}\rangle, \quad (594)$$

with $\hat{L}_{QQ}$ denoting the time-independent projected transition operator $\hat{L}_{QQ} = \hat{Q}\hat{L}\hat{Q}$. The projection operator $\hat{Q}$ is the complement of the time-independent projection operator $\hat{P}$ defined by Eq. (587). In Eqs. (593) and (594), the tilde indicates the usual Kubo transform for a system in thermal equilibrium.

From the above, we see that the nonlinear, non-Markovian generalized phenomenological equations (588a) and (588b) become linear, non-Markovian equations (592), in the regime close to thermal equilibrium. In the Markovian approximation, the linear, non-Markovian equations assume a form that is identical to the linear phenomenological equations given by Eq. (508) from the linear domain of QSM theory for the local densities $\{\langle\hat{\rho}_j(\vec{r},t)\rangle\}$ of conserved quantities, except that the phenomenological time derivative is replaced by an instantaneous time derivative.

As for Eq. (508), the phenomenological coefficient $M_{jk}(\vec{r},\vec{r}\,')$ resulting from the Markovian approximation of Eq. (592) for the case of local densities $\{\langle\hat{\rho}_j(\vec{r},t)\rangle\}$ of conserved quantities is given by Eq. (505), with $M^S_{jk}(\vec{r},\vec{r}')$ given by Eq. (506). However, $M^C_{jk}(\vec{r},\vec{r}')$ is given by

$$M^C_{jk}(\vec{r},\vec{r}') = \lim_{z\to 0^+}\left\langle \dot{\rho}^\dagger_j(\vec{r})\middle|\hat{Q}(z\,\hat{Q}+\hat{L}_{QQ})^{-1}\hat{Q}\,\middle|\tilde{\rho}_k(\vec{r}')\rho_{\rm eq}\right\rangle \tag{595}$$

rather than Eq. (507).

Fluctuations $\delta\hat{O}_j(\vec{r},\,t)=\hat{O}_j(\vec{r},\,t)-\langle\hat{O}_j(\vec{r},\,t)\rangle$ about the mean path, defined by $\langle\hat{O}_j(\vec{r},\,t)\rangle$, can be described by the vectors $\langle\delta O^\dagger_{j,t}(\vec{r}\,)|=\langle O^\dagger_{j,t}(\vec{r}\,)|-\langle\hat{O}_j(\vec{r},\,t)\rangle\langle I|$. The time derivative of these vectors is given by $(\partial/\partial t)\langle\delta O^\dagger_{j,t}(\vec{r}\,)|=\langle\dot{O}^\dagger_{j,t}(\vec{r}\,)|-\langle\dot{\hat{O}}_j(\vec{r},\,t)\rangle\langle I|$. Application of the projection operators $\hat{P}(t)$ and $\hat{Q}(t)$ to the latter equation gives

$$\frac{\partial}{\partial t}\langle\delta O^\dagger_{j,t}(\vec{r})| = -\sum_k\left[i\int_{\mathcal{D}}dV'\Omega_{jk}(\vec{r},\vec{r}';t)\langle\delta O^\dagger_{k,t}(\vec{r}')| + \int_{\mathcal{D}}dV'\int_0^t dt'\mathbb{K}_{jk}(\vec{r}t,\vec{r}'t')\langle\delta O^\dagger_{k,t'}(\vec{r}')|\right] + \langle f_{j,t}(\vec{r})|, \tag{596}$$

where

$$i\Omega_{jk}(\vec{r},\vec{r}';t)=\sum_l\int_{\mathcal{D}}dV''\mathbb{M}^S_{jl}(\vec{r},\vec{r}'';t)\chi^{-1}_{lk}(\vec{r}'',\vec{r}';t), \tag{597}$$

$$\mathbb{K}_{jk}(\vec{r}t,\vec{r}'t')=\sum_l\int_{\mathcal{D}}dV''\mathbb{M}^C_{jl}(\vec{r}t,\vec{r}''t')\chi^{-1}_{lk}(\vec{r}'',\vec{r}',t'), \tag{598}$$

and

$$\langle f_{j,t}(\vec{r})| = -\langle\delta_t O^\dagger_j(\vec{r})|\hat{L}\,\hat{Q}(t)\hat{\Phi}_{QQ}(t,0) - \int_0^t dt'\langle\delta_t O^\dagger_j(\vec{r})|\hat{L}\,\hat{Q}(t)\hat{\Phi}_{QQ}(t,t')\hat{Q}(t') \times\exp(-\hat{L}t'), \tag{599}$$

with

$$\mathbb{M}^S_{jl}(\vec{r},\vec{r}'';t)=\left\langle\delta_t O^\dagger_j(\vec{r})\middle|\hat{L}\,\middle|\delta_t\tilde{O}_l(\vec{r}'')\bar{\rho}_t\right\rangle \tag{600}$$

and

$$\mathbb{M}^C_{jl}(\vec{r}t,\vec{r}''t') = -\langle\delta_t O^\dagger_j(\vec{r})|\hat{L}\,\hat{Q}(t)\hat{\Phi}_{QQ}(t,t') \times\hat{Q}(t')\hat{L}\,|\delta_{t'}\tilde{O}_l(\vec{r}'')\bar{\rho}_{t'}\rangle. \tag{601}$$

The generalized Kubo transform in $|\delta_t\tilde{O}_l(\vec{r}'')\bar{\rho}_t\rangle$ $[|\delta_{t'}\tilde{O}_l(\vec{r}'')\bar{\rho}_{t'}\rangle]$ depends on $t[t']$.

The equation of motion given by Eq. (596) is an exact generalized Langevin equation for the vectors $\langle\delta O^\dagger_{j,t}(\vec{r}\,)|$ corresponding to the fluctuations $\delta\hat{O}_j(\vec{r},\,t)=\hat{O}_j(\vec{r},\,t)-\langle\hat{O}_j(\vec{r},\,t)\rangle$ about the mean path. It represents a spatially dependent generalization of the generalized Langevin equation given by Eq. (574). If the fluctuations $\delta\hat{O}_j(\vec{r},\,t)$ are about the state of thermal equilibrium, Eq. (596) reduces to a spatially dependent generalization of the generalized Langevin equation given by Eq. (554) in our discussion of the Mori–Zwanzig projection operator formalism.

With Eq. (596) at our disposal, we can consider the time evolution of the displacements $\langle\delta\hat{O}_j(\vec{r},\,t)\rangle=\langle\hat{O}_j(\vec{r},\,t)\rangle-\langle\hat{O}_j(\vec{r}\,)\rangle_{\rm eq}$ from thermal equilibrium. Making use of Eq. (596) to form the time derivative $(\partial/\partial t)\langle\delta O^\dagger_{j,t}(\vec{r}\,)|\rho_{\rm eq}\rangle$ in the expression $(\partial/\partial t)\langle\delta\hat{O}_j(\vec{r},\,t)\rangle = -(\partial/\partial t)\langle\delta O^\dagger_{j,t}(\vec{r}\,)|\rho_{\rm eq}\rangle$, we obtain

$$\frac{\partial}{\partial t}\langle\delta\hat{O}_j(\vec{r},t)\rangle = -\sum_k\left[i\int_{\mathcal{D}}dV'\Omega_{jk}(\vec{r},\vec{r}';t) \times\langle\delta\hat{O}_k(\vec{r}',t)\rangle + \int_{\mathcal{D}}dV'\int_0^t dt'\mathbb{K}_{jk} \times(\vec{r}\,t,\vec{r}'t')\langle\delta\hat{O}_k(\vec{r}',t')\rangle\right] + \mathbb{I}_j(\vec{r},t), \tag{602}$$

where

$$\mathbb{I}_j(\vec{r},t) = \left\langle\delta_t O^\dagger_j(\vec{r})\middle|\hat{L}\,\hat{Q}(t)\hat{\Phi}_{QQ}(t,0)\middle|\rho_{\rm eq}\right\rangle + \int_0^t dt'\left\langle\delta_t O^\dagger_j(\vec{r})\middle|\hat{L}\,\hat{Q}(t)\hat{\Phi}_{QQ}(t,t')\hat{Q}(t')\middle|\rho_{\rm eq}\right\rangle. \tag{603}$$

The exact result given by Eq. (602) represents a spatially dependent generalization of the equation of motion given by Eq. (580) for the evolution of the displacements from thermal equilibrium. If the system is close to thermal equilibrium, Eq. (602) assumes the form of a spatially dependent generalization of the evolution equation given by Eq. (548) in our discussion of the the Mori–Zwanzig projection operator formalism.

Let us suppose that the thermodynamic coordinates $\{\langle\hat{O}_j(\vec{r},t)\rangle\}$ correspond to the average local densities $\{\langle\hat{\rho}_j(\vec{r},t)\rangle\}$ of conserved quantities. For this case, one can demonstrate that the equations of motion given by Eq. (588a) can be cast in the form

$$\langle\hat{\vec{J}}_j(\vec{r},\,t)\rangle = \langle\hat{\vec{J}}_j(\vec{r})\rangle_{\bar{\rho}(t)} + \sum_k\int_{\mathcal{D}}dV' \times\int_0^t dt'\,\vec{\vec{\mathfrak{M}}}^C_{jk}(\vec{r}\,t,\vec{r}'t')\cdot\vec{\nabla}\Lambda_k(\vec{r}',t'), \tag{604}$$

where

$$\langle\hat{\vec{J}}_j(\vec{r})\rangle_{\bar{\rho}(t)}=\left\langle\vec{J}^\dagger_j(\vec{r})\middle|\bar{\rho}_t\right\rangle \tag{605}$$

and

$$\vec{\vec{\mathfrak{M}}}^C_{jk}(\vec{r}\,t,\vec{r}'t')=\left\langle\vec{J}^\dagger_j(\vec{r})\middle|\hat{Q}(t)\hat{\Phi}_{QQ}(t,t')\hat{Q}(t')\middle|\vec{\tilde{J}}_k(\vec{r}')\bar{\rho}_{t'}\right\rangle. \tag{606}$$

The generalized Kubo transform in $|\vec{\tilde{J}}_k(\vec{r}')\bar{\rho}_{t'}\rangle$ depends on t'. Equation (604) represents a spatially and temporally nonlocal generalization of the linear phenomenological equations for the current densities in nonequilibrium thermodynamics for the case of continuous systems.

Generalized phenomenological equations identical in form to Eq. (604) were first established by Robertson using the aforementioned defective projection operator. Unlike Robertson's generalized phenomenological equations, the result given by Eq. (604) conforms to the conservation of probability requirement $\langle I \mid \rho_t \rangle = \langle I \mid \bar{\rho}_t \rangle = 1$.

If the system is close to thermal equilibrium, we can use Eq. (592) to construct the following approximate expression for the average current densities:

$$\langle \hat{\vec{J}}_j(\vec{r},t)\rangle = \sum_k \left[\int_{\mathcal{D}} dV' \, \vec{\mathfrak{M}}^S_{jk}(\vec{r},\vec{r}')\Lambda_k(\vec{r}',t) + \int_{\mathcal{D}} dV' \times \int_0^t dt' \, \vec{\vec{\mathfrak{M}}}^C_{jk}(\vec{r},\vec{r}';t-t')\cdot\vec{\nabla}'\Lambda_k(\vec{r}',t') \right], \tag{607}$$

where

$$\vec{\mathfrak{M}}^S_{jk}(\vec{r},\vec{r}\,') = \begin{cases} -\langle \vec{J}^\dagger_j(\vec{r}) | \tilde{O}_k(\vec{r}')\rho_{\text{eq}}\rangle \\ -\langle O^\dagger_k(\vec{r}') | \vec{\tilde{J}}_j(\vec{r})\rho_{\text{eq}}\rangle \end{cases} \tag{608}$$

and

$$\vec{\vec{\mathfrak{M}}}^C_{jk}(\vec{r},\vec{r}';t-t') = \langle \vec{J}^\dagger_j(\vec{r}) | \hat{Q}\exp[-\hat{L}_{QQ}(t-t')]\hat{Q} \, | \vec{\tilde{J}}_k(\vec{r}')\rho_{\text{eq}}\rangle. \tag{609}$$

In the Markovian approximation, Eq. (607) assumes a form that is identical to the form given by Eq. (509) from the linear domain of QSM theory. The quantity $\vec{\mathfrak{M}}^S_{jk}(\vec{r},\vec{r}')$ appearing in the Markovian approximation of Eq. (607) is identical to the same quantity appearing in Eq. (509). However, $\vec{\vec{\mathfrak{M}}}^C_{jk}(\vec{r},\vec{r}')$ is given by

$$\vec{\vec{\mathfrak{M}}}^C_{jk}(\vec{r},\vec{r}\,') = \lim_{z\to 0^+} \langle \vec{J}^\dagger_j(\vec{r}) | \hat{Q}(z\,\hat{Q} + \hat{L}_{QQ})^{-1}\hat{Q} \, | \vec{\tilde{J}}_k(\vec{r}')\rho_{\text{eq}}\rangle \tag{610}$$

rather than Eq. (511) or (512).

In order to work with the various equations of motion discussed above, one must be able to determine the thermodynamic parameters $\{\Lambda_j(\vec{r},\,t)\}$. These quantities can be determined by solving the set of equations (432), which is equivalent to solving the basic equations $\langle \hat{O}_j(\vec{r},\,t)\rangle = \text{Tr}\;\hat{\rho}(t)\hat{O}_j(\vec{r}\,)$ defining the thermodynamic state of the system, where $\hat{\rho}(t)$ is given by Eq. (425).

Substituting Eq. (588b) into Eq. (452), we obtain the following closed set of nonlinear, non-Markovian integral equations of motion for the thermodynamic parameters:

$$\frac{\partial}{\partial t}\Lambda_j(\vec{r},t) = -\sum_k \left[\int_{\mathcal{D}} dV' \mathfrak{N}^S_{jk}(\vec{r},\vec{r}';t)\Lambda_k(\vec{r}',t) + \int_{\mathcal{D}} dV' \int_0^t dt' \mathfrak{N}^C_{jk}(\vec{r}\,t,\vec{r}'t')\Lambda_k(\vec{r}',t') \right], \tag{611}$$

where

$$\mathfrak{N}^S_{jk}(\vec{r},\vec{r}';t) = \sum_l \int_{\mathcal{D}} dV'' \, \chi^{-1}_{jl}(\vec{r},\vec{r}'';t)M^S_{lk}(\vec{r}'',\vec{r}';t) \tag{612}$$

and

$$\mathfrak{N}^C_{jk}(\vec{r}\,t,\vec{r}'t') = \sum_l \int_{\mathcal{D}} dV'' \chi^{-1}_{jl}(\vec{r},\vec{r}'';t)M^C_{lk}(\vec{r}''t,\vec{r}'t'). \tag{613}$$

The result given by Eq. (611) represents a spatially dependent generalization of the equation of motion given by Eq. (582) for the evolution of spatially independent thermodynamic parameters.

From our considerations of both spatially independent and spatially dependent thermodynamic corrdinates, we find that the maximum entropy approach to nonequilibrium processes enables us to determine all of the quantities appearing in nonequilibrium thermodynamics. Moreover, this approach to nonequilibrium processes is consistent with the general evolution theorem given by Eq. (367), which was established on the basis of the time-dependent version of the maximum entropy principle. As discussed in Sections 5.3 and 5.4, this evolution theorem plays a fundamental role in nonequilibrium thermodynamics.

E. Dual Lanczos Transformation Theory

The character of a dynamical system is usually described in terms of only a few spectral and temporal properties. One would like to determine these properties without resorting to the difficult, if not impossible, task of solving the equations of motion for the system. This goal may be achieved, or approximately achieved, by utilizing the formal apparatus of the author's dual Lanczos transformation theory, which provides a number of universal formulas and other results for handling such problems. Unlike more traditional approaches to spectral and temporal problems, dual Lanczos transformation theory is written in such a way that the same formal apparatus applies to classical and quantum systems, regardless of whether the underlying dynamics is reversible or irreversible.

The central concept in dual Lanczos tranformation theory is the idea of determining spectral and temporal properties by means of the extraction and utilization of dynamically embedded information. In essence, the implementation of this idea entails the following basic steps:

(i) Build a dual Lanczos vector space embedded with the relevant dynamical information, (ii) extract the information embedded in the dual Lanczos vector space, and (iii) utilize the extracted information to determine the spectral and temporal properties of interest.

Apart from determining the spectral and temporal properties of interest, one can perform suitable mathematical operations on dynamically embedded information to (i) construct reduced or contracted equations of motion, (ii) investigate generalized Langevin equations and memory function hierarchies, (iii) make rigorous statements about expected temporal character, including statements about the possible existence of separations in time scales and the role of memory effects, and (iv) investigate other interesting properties that might arise during the implementation of the theory.

An especially appealing feature of dual Lanczos transformation theory its algebraic character. This feature has enabled the author to exploit developments in symbolic processing to put the theory in symbolic code for the purpose of using a computer to obtain exact and/or approximate analytic solutions to spectral and temporal problems. Symbolic codes may be used to carry out most of the required operations with minimal manual effort and computer time. The capability of doing this has afforded us with the opportunity to explore a variety of model systems and to connect their spectral and temporal properties to microscopic interactions (and phenomenological parameters in mesoscopic descriptions) through analytic formulas in a physically and mathematically transparent way. The utility of dual Lanczos transformation theory in the analytic treatment of spectral and temporal properties has been illustrated in a number of theoretical studies including light/neutron scattering, resonant γ-ray absorption, spectral and temporal properties of the Brownian harmonic oscillator, nonradiative decay, excited state dynamics, collision-induced light absorption, and harmonic generation in laser-driven systems.

1. Basic Structure of System Dynamics

In order to completely characterize the dynamics of a system, we must introduce a system propagator $\hat{U}(t)$ that bears information about the time-reversal properties of the system by including both its forward and backward time evolution. The usual propagator $\exp(-\hat{L}t)$ is insufficient when the transition operator $\hat{L} = i\hat{\mathcal{L}}$ possesses broken time-reversal symmetry, i.e., when $\hat{\tilde{L}} \neq -\hat{L}$, where $\hat{\tilde{L}}$ is the time-reversed form of $\hat{L}$. Of course, the usual propagator $\exp(-\hat{L})$ is sufficient when we are dealing with a reversible system, i.e., when the symmetry relation $\hat{\tilde{L}} = -\hat{L}$ holds.

On the basis of time-reversal arguments, one can assert that the backward time evolution of a system is given by $\theta(-t)\exp(\hat{\tilde{L}}t)$. Hence, the propagator $\hat{U}(t)$ assumes the form

$$\hat{U}(t) = \hat{U}^{>}(t) + \hat{U}^{<}(t), \tag{614}$$

where

$$\hat{U}^{>}(t) = \theta(t)\exp(-\hat{L}t) \tag{615}$$

and

$$\hat{U}^{<}(t) = \theta(-t)\exp(\hat{\tilde{L}}t), \tag{616}$$

with $\theta(t)$ denoting the Heaviside step function. The operators $\hat{U}^{>}(t)$ and $\hat{U}^{<}(t)$, respectively, describe the retarded (forward in time) and advanced (backward in time) dynamics of a system.

The propagator $\hat{U}(t)$ defined by Eqs. (614)–(616) possesses the following properties: (i) $\hat{U}(t) = \hat{U}^{>}(t)$ for $t > 0$ and $\hat{U}(t) = \hat{U}^{<}(t)$ for $t < 0$. (ii) $\lim_{t\to 0^+}\hat{U}(t) = \lim_{t\to 0^+}\hat{U}^{>}(t) = \hat{I}$ and $\lim_{t\to 0^-}\hat{U}(t) = \lim_{t\to 0^-}\hat{U}^{<}(t) = \hat{I}$. (iii) $\hat{U}(t)$ is invariant with respect to the time-reversal transformation $\hat{L}, t \to \hat{\tilde{L}}, -t$.

It should be noted that $\hat{U}^{<}(t) = \theta(-t)\exp(-\hat{L}t)$ when the symmetry relation $\hat{\tilde{L}} = -\hat{L}$ holds. Hence, the propagator $\hat{U}(t)$ assumes the usual form $\hat{U}(t) = \exp(-\hat{L}t)$ for the case of reversible systems.

2. Evolution of Dynamical Vectors

In our discussion of dual Lanczos transformation theory, we shall define the inner product $(A|B)$ of the dynamical vectors $(A|$ and $|B)$ as $(A|B) = \mathrm{Tr}\,\hat{A}\hat{B}$ for quantum systems and $(A|B) = \mathrm{Tr}\,AB$ for classical systems, where the quantum and classical traces are to be interpreted in the same fashion as discussed earlier. The motivation for adopting this definition is to render the mathematics to be such that the treatment of the properties of a physical system is independent of the nature of the underlying dynamics. The formal connection between the vectors $(A|$ and $|B)$ and the vectors used in our earlier discussions of dynamical systems is given by $(A| = \langle A^*|$ and $|B) = |B\rangle$ for classical systems and $(A| = \langle A^\dagger|$ and $|B) = |B\rangle$ for quantum systems.

The dynamical vector $|\rho_t)$ corresponding to the density operator or probability density for a system is given by $|\rho_t) = \hat{U}(t)|\rho_0)$ for all t. Introducing the propagator $\hat{U}(t)$ defined by Eq. (614), $|\rho_t)$ may be resolved into retarded and advanced components, i.e., $|\rho_t) = |\rho_t^{>}) + |\rho_t^{<})$, where $|\rho_t) = |\rho_t^{>})$ for $t > 0$ and $|\rho_t) = |\rho_t^{<})$ for $t < 0$, with $|\rho_t^{>}) = \hat{U}^{>}(t)\,|\rho_0^{>})$ and $|\rho_t^{<}) = \hat{U}^{<}(t)\,|\rho_0^{<})$. The vectors $|\rho_t)$, $|\rho_t^{>})$, and $|\rho_t^{<})$ satisfy the initial conditions $\lim_{t\to 0^+}|\rho_t) = \lim_{t\to 0^+}|\rho_t^{>}) = |\rho_0^{>})$ and $\lim_{t\to 0^-}|\rho_t) = \lim_{t\to 0^-}|\rho_t^{<}) = |\rho_0^{<})$, where $|\rho_0^{>}) = |\rho_0^{>}) = |\rho_0)$.

The dynamical vector $(A_t|$ corresponding to a classical or quantum dynamical variable is given by $(A_t| = (A_0|\,\hat{U}(t)$ for all t. As with $|\rho_t)$, we can resolve $(A_t|$ into retarded and advanced components, i.e., $(A_t| = (A_t^{>}| + (A_t^{<}|$, where $(A_t| = (A_t^{>}|$ for $t > 0$ and $(A_t| = (A_t^{<}|$ for $t < 0$, with $(A_t^{>}| = (A_0^{>}|\,\hat{U}^{>}(t)$ and $(A_t^{<}| = (A_0^{<}|\,\hat{U}^{<}(t)$. The vectors $(A_t|$, $(A_t^{>}|$, and $(A_t^{<}|$ satisfy the initial conditions $\lim_{t\to 0^+}(A_t| = \lim_{t\to 0^+}(A_t^{>}| = (A_0^{>}|$ and $\lim_{t\to 0^-}(A_t| = \lim_{t\to 0^-}(A_t^{<}| = (A_0^{<}|$, where $(A_0^{>}| = (A_0^{<}| = (A_0|$.

3. Decomposition of System Dynamics

All of the dynamical information about a system may be thought to be embedded in some abstract space $\mathcal{G}$. In general, it is very difficult to utilize this information to determine the small number of spectral and temporal properties of interest. In part, this difficulty stems from the problem of eliminating the irrelevant information in an efficient way and the lack of adequate analytical and numerical tools for dealing with the relevant information. The work of Mori and Zwanzig suggests that dynamical information may be dealt with by using orthogonal projection operators.

Zwanzig used orthogonal projection operators to decompose the dynamical vector $|\rho_t)$ into relevant and irrelevant parts for the forward direction of time. With this decomposition, Zwanzig obtained the non-Markovian master equation given by Eq. (546) for the relevant part $|\rho_t^{\mathrm{R}})$ of $|\rho_t)$. As indicated earlier, the dynamical vector $|\rho_t^{\mathrm{R}})$ bears sufficient information for determining the properties of interest in the forward direction of time.

Mori introduced a decomposition of the dynamics into relevant and irrelevant parts by the application of orthogonal projection operators directly to the vector equation of motion for dynamical variables. As discussed earlier, this approach leads to Mori's generalized Langevin equation. [See Eq. (554).]

In dual Lanczos transformation theory, orthogonal projection operators are used to decompose the dynamics of a system into relevant and irrelevant parts in the same spirit as in the Mori–Zwanzig projection operator formalism, but with a differing motivation and decomposition. More specifically, orthogonal projection operators are used to decompose the retarded or advanced dynamics into relevant and irrelevant parts that are completely decoupled. The subdynamics for each of these parts is an independent and closed subdynamics of the system dynamics. With this decomposition, we are able to completely discard the irrelevant information and focus our attention solely on the closed subdynamics of the relevant part.

Consider a projection operator $\hat{P}$ that projects onto a subspace $\mathcal{G}_P^{>}$ that is embedded with information about certain spectral and temporal properties of a system in the forward direction of time. The operator $\hat{Q} = \hat{I} - \hat{P}$ is the orthogonal complement of $\hat{P}$ and projects onto the subspace $\mathcal{G}_Q^{>}$. The abstract space $\mathcal{G}^{>}$ embedded with all of the dynamical information about a system in the forward direction of time is simply the sum $\mathcal{G}_P^{>} + \mathcal{G}_Q^{>}$ of the subspaces. The orthogonal projection operators $\hat{P}$ and $\hat{Q}$ are assumed to satisfy the usual relations $\hat{P} + \hat{Q} = \hat{I}$, $\hat{P}^2 = \hat{P}$, $\hat{Q}^2 = \hat{Q}$, and $\hat{P}\hat{Q} = \hat{Q}\hat{P} = \hat{0}$, where $\hat{0}$ is the null operator.

Introducing the projection operators $\hat{P}$ and $\hat{Q}$, we decompose the retarded propagator $\hat{U}^{>}(t)$ as follows:

$$\hat{U}{>}(t) = \hat{U}_{PP}^{>}(t) + \hat{U}_{PQ}^{>}(t) + \hat{U}_{QP}^{>}(t) + \hat{U}_{QQ}^{>}(t). \quad (617)$$

The projected propagators $\hat{U}_{PP}^{>}(t) = \theta(t)\,\hat{P}\exp(-\hat{L}t)\hat{P}$ and $\hat{U}_{QQ}^{>}(t) = \theta(t)\hat{Q}\exp(-\hat{L}t)\hat{Q}$ describe the retarded subdynamics of the orthogonal subspaces $\mathcal{G}_P^{>}$ and $\mathcal{G}_Q^{>}$, respectively. The cross terms $\hat{U}_{PQ}^{>}(t) = \theta(t)\hat{P}\exp(-\hat{L}t)\hat{Q}$ and $\hat{U}_{QP}^{>}(t) = \theta(t)\,\hat{Q}\exp(-\hat{L}t)\hat{P}$ represent the interplay between the two subdynamics due to their interaction through $\hat{Q}\hat{L}\hat{P}$ and $\hat{P}\hat{L}\hat{Q}$.

One can demonstrate that the subdynamics of $\mathcal{G}_P^{>}$ and $\mathcal{G}_Q^{>}$ are described by the equations of motion

$$\frac{d}{dt}\hat{U}_{PP}^{>}(t) = -\hat{L}_{PP}\hat{U}_{PP}^{>}(t) + \int_0^t dt'\,\{\hat{L}_{PQ} \times \exp[-\hat{L}_{QQ}(t-t')]\hat{L}_{QP}\}\,\hat{U}_{PP}^{>}(t') \quad (618)$$

and

$$\frac{d}{dt}\hat{U}_{QQ}^{>}(t) = -\hat{L}_{QQ}\hat{U}_{QQ}^{>}(t) + \int_0^t dt'\,\{\hat{L}_{QP} \times \exp[-\hat{L}_{PP}(t-t')]\hat{L}_{PQ}\}\,\hat{U}_{QQ}^{>}(t') \quad (619)$$

for $t > 0$, where $\hat{L}_{PP} = \hat{P}\hat{L}\hat{P}$, $\hat{L}_{PQ} = \hat{P}\hat{L}\hat{Q}$, $\hat{L}_{QP} = \hat{Q}\hat{L}\hat{P}$, and $\hat{L}_{QQ} = \hat{Q}\hat{L}\hat{Q}$.

The first term on the right side of Eq. (618) represents the influence of the intrinsic subdynamics of $\mathcal{G}_P^{>}$, while the second term represents the influence of the subdynamics of $\mathcal{G}_Q^{>}$ through the interactions $\hat{L}_{PQ}$ and $\hat{L}_{QP}$. A similar interpretation may be given to Eq. (619).

Assuming that $\hat{P}$ projects onto a dynamically invariant subspace of $\mathcal{G}^{>}$, the subdynamics of $\mathcal{G}_P^{>}$ and $\mathcal{G}_Q^{>}$ are completely decoupled due to the elimination of the interactions $\hat{L}_{PQ}$ and $\hat{L}_{QP}$. This decoupling is embodied in the following simplified forms of Eqs. (618) and (619):

$$\frac{d}{dt}\hat{U}_{PP}^{>}(t) = -\hat{L}_{PP}\hat{U}_{PP}^{>}(t) \quad (620)$$

and

$$\frac{d}{dt}\hat{U}^{>}_{QQ}(t) = -\hat{L}_{QQ}\hat{U}^{>}_{QQ}(t) \tag{621}$$

for $t > 0$.

In essence, the use of a projection operator $\hat{P}$ that projects onto a dynamically invariant subspace of $\mathcal{G}^{>}$ enables one to decompose the retarded dynamics of a system into two completely independent pieces:

$$\hat{U}^{>}(t) = \hat{U}^{>}_{PP}(t) + \hat{U}^{>}_{QQ}(t), \tag{622}$$

where $\hat{U}^{>}_{PP}(t) = \theta(t)\exp(-\hat{L}_{PP}t)$ and $\hat{U}^{>}_{QQ}(t) = \theta(t)\exp(-\hat{L}_{QQ}t)$ describe the closed subdynamics of $\mathcal{G}^{>}_{P}$ and $\mathcal{G}^{>}_{Q}$, respectively.

A decomposition similar to Eq. (622) for the advanced dynamics of a system may be accomplished by using a projection operation $\hat{P}'$ that projects onto a dynamically invariant subspace $\mathcal{G}^{<}_{P'}$. This leads to

$$\hat{U}^{<}(t) = \hat{U}^{<}_{P'P'}(t) + \hat{U}^{<}_{Q'Q'}(t), \tag{623}$$

where $\hat{U}^{<}_{P'P'}(t) = \theta(-t)\exp(\hat{\tilde{L}}_{P'P'}t)$ and $\hat{U}^{<}_{Q'Q'}(t) = \theta(-t)\exp(\hat{\tilde{L}}_{Q'Q'}t)$, with $\hat{\tilde{L}}_{P'P'} = \hat{P}'\hat{\tilde{L}}\hat{P}'$ and $\hat{\tilde{L}}_{Q'Q'} = \hat{Q}'\hat{\tilde{L}}\hat{Q}'$.

The dynamically invariant subspaces $\mathcal{G}^{>}_{P}$ and $\mathcal{G}^{>}_{P'}$ are equivalent when the dynamics satisfies the symmetry relation $\hat{\tilde{L}} = \pm\hat{L}$, as in the case of reversible systems ($\hat{\tilde{L}} = -\hat{L}$), provided the projection operators $\hat{P}$ and $\hat{P}'$ are equivalent. Otherwise, $\mathcal{G}^{>}_{P}$ and $\mathcal{G}^{<}_{P'}$ are different.

In dual Lanczos transformation theory, we project onto a dynamically invariant subspace $\mathcal{G}^{>}_{\mathrm{LZ}}$ or $\mathcal{G}^{<}_{\mathrm{LZ}}$ called a dual Lanczos vector space. This projection is accomplished with the projection operator

$$\hat{I}^{\mathrm{LZ}} = \sum_{s} |p_s)(r_s|, \tag{624}$$

which represents the identity operator for the dual Lanczos vector space employed. The index s runs over the dual Lanczos vectors $\{(r_s|\}$ and $\{|p_s)\}$. These vectors form a biorthonormal basis satisfying the orthonormality relation $(r_s \mid p_{s'}) = \delta_{s,s'}$. Provided we have properly biased the information embedded in the dual Lanczos vector space, we can completely discard the irrelevant subdynamics of the orthogonal subspace and work only with the subdynamics of the dual Lanczos vector space.

4. Dual Lanczos Transformations

Given the starting vectors $(r_0|$ and $|p_0)$, the remaining basis vectors for the dual Lanczos vector space $\mathcal{G}^{>}_{\mathrm{LZ}}$ may be generated by means of the recursion relations

$$\beta_{s+1}|p_{s+1}) = (\hat{L} - \alpha_s\hat{I})|p_s) - \beta_s|p_{s-1}) \tag{625}$$

and

$$\beta_{s+1}(r_{s+1}| = (r_s|(\hat{L} - \alpha_s\hat{I}) - \beta_s(r_{s-1}|, \tag{626}$$

where $s \geq 0$, and $(r_{-1}|$ and $|p_{-1})$ are null vectors.

The Lanczos parameters $\{\alpha_s\}$ and $\{\beta_{s+1}\}$ appearing in Eqs. (625) and (626) are the only nonvanishing matrix elements of the transition operator $\hat{L}$ in the dual Lanczos representation:

$$\alpha_s = (r_s|\hat{L}|p_s) \tag{627}$$

and

$$\beta_{s+1} = (r_s|\hat{L}|p_{s+1}) \tag{628a}$$

$$= (r_{s+1}|\hat{L}|p_s). \tag{628b}$$

The dual Lanczos vector space generated with the starting vectors $(r_0|$ and $|p_0)$ is finite dimensional when the Lanczos parameter β_N vanishes for some finite value of N. For such cases, the dual Lanczos vector space is a dynamically invariant subspace of dimensionality N. This situation is realized when the operation $(r_N|\hat{L}$ or $\hat{L}|p_N)$ does not generate any new dynamical information, i.e., information not already contained in the vectors $\{(r_s|; s = 0, \ldots, N-1\}$ and $\{|p_s); s = 0, \ldots, N-1\}$. One should bear in mind that a dual Lanczos vector space is a dynamically invariant subspace by construction, regardless of its dimensionality.

The basis vectors for the dual Lanczos vector space $\mathcal{G}^{<}_{\mathrm{LZ}}$ may be generated with recursion relations identical in form to Eqs. (625) and (626). One only needs to replace $\hat{L}$ with $\hat{\tilde{L}}$. As indicated earlier, the space $\mathcal{G}^{<}_{\mathrm{LZ}}$ does not have to be generated when $\hat{\tilde{L}} = \pm\hat{L}$. Even for cases not satisfying this symmetry relation, we have not encountered a problem for which it is necessary to actually build $\mathcal{G}^{<}_{\mathrm{LZ}}$. Hence, we shall confine the remainder of our discussion to the dual Lanczos vector space $\mathcal{G}^{>}_{\mathrm{LZ}}$.

Equations (625) and (626) can be written as matrix equations:

$$\mathbf{L}\mathbf{X}^{P} = \mathbf{X}^{P}\mathbf{L}^{\mathrm{LZ}}_{0} \tag{629}$$

and

$$\mathbf{X}^{R}\mathbf{L} = \mathbf{L}^{\mathrm{LZ}}_{0}\mathbf{X}^{R}, \tag{630}$$

where $\mathbf{X}^{P}$ $[\mathbf{X}^{R}]$ is the matrix formed by the components of the right-hand Lanczos vectors $\{|p_s)\}$ [left-hand Lanczos vectors $\{(r_s|\}$] and

$$\mathbf{L}_0^{\mathrm{LZ}} = \begin{bmatrix} \alpha_0 & \beta_1 & & \\ \beta_1 & \alpha_1 & \beta_2 & \\ & \beta_2 & \alpha_2 & \ddots \\ & & \ddots & \ddots \end{bmatrix} \tag{631}$$

is the complex symmetric tridiagonal matrix formed by $\hat{L}$ in the dual Lanczos representation.

Since the dual Lanczos vectors are normalized to unity $[(r_s|p_{s'}) = \delta_{s,s'}]$,

$$\mathbf{X}^R\mathbf{X}^P = \mathbf{I}^{\mathrm{LZ}}, \tag{632}$$

where $\mathbf{I}^{\mathrm{LZ}}$ is a unit matrix, with the number of rows and columns of $\mathbf{I}^{\mathrm{LZ}}$ equal to the dimensionality of the dual Lanczos vector space. By definition, $\mathbf{X}^P[\mathbf{X}^R]$ is the right [left] inverse of $\mathbf{X}^R[\mathbf{X}^P]$. In general, $\mathbf{X}^P[\mathbf{X}^R]$ is not the left [right] inverse of $\mathbf{X}^R[\mathbf{X}^P]$.

It follows from Eq. (632) that Eqs. (629) and (630) can be written as a single matrix equation:

$$\mathbf{X}^R\mathbf{L}\mathbf{X}^P = \mathbf{L}_0{}^{\mathrm{LZ}}. \tag{633}$$

This result represents a general transformation that transforms the matrix $\mathbf{L}$, regardless of its symmetry, into a complex symmetric tridiagonal matrix $\mathbf{L}_0^{\mathrm{LZ}}$. Such transformations are called dual Lanczos transformations.

We can interpret Eq. (633) as the matrix form of the operator equation

$$\hat{X}^R\hat{L}\hat{X}^P = \hat{L}_0^{\mathrm{LZ}}, \tag{634}$$

where

$$\hat{X}^R\hat{X}^P = \hat{I}^{\mathrm{LZ}}, \tag{635}$$

with $\hat{I}^{\mathrm{LZ}}$ denoting the identity operator for the dual Lanczos vector space. Equation (634) is simply an operator version of a dual Lanczos transformation. It represents the transformation of the operator $\hat{L}$ into the operator $\hat{L}_0^{\mathrm{LZ}}$ obtained by projecting $\hat{L}$ onto a dual Lanczos vector space, i.e., $\hat{L}_0^{\mathrm{LZ}} = \hat{I}^{\mathrm{LZ}}\hat{L}\hat{I}^{\mathrm{LZ}}$.

More generally, we can write a dual Lanczos transformation as

$$\hat{X}^R\hat{f}(\hat{L})\hat{X}^P = \hat{f}(\hat{X}^R\hat{L}\hat{X}^P) \tag{636a}$$

$$= \hat{f}\big(\hat{L}_0^{\mathrm{LZ}}\big), \tag{636b}$$

where $\hat{f}(\hat{L})$ is any operator function of $\hat{L}$. Unless $\hat{X}^P[\hat{X}^R]$ is the left [right] inverse of $\hat{X}^R[\hat{X}^P]$, $\hat{f}(\hat{L}) \neq \hat{X}^P\hat{f}(\hat{L}_0^{\mathrm{LZ}})\hat{X}^R$.

It follows from Eq. (636b) and the fact that a dual Lanczos vector space is a dynamically invariant subspace by construction that the projection of the retarded dynamics of a system onto the closed subdynamics of a dual Lanczos vector space may be accomplished with a dual Lanczos transformation. More specifically,

$$\theta(t)\,\hat{I}^{\mathrm{LZ}}\exp(-\hat{L}t)\hat{I}^{\mathrm{LZ}} = \theta(t)\,\hat{X}^R\exp(-\hat{L}t)\hat{X}^P \tag{637a}$$

$$= \theta(t)\,\exp\big(-\hat{L}_0^{\mathrm{LZ}}t\big). \tag{637b}$$

In general, a dual Lanczos vector space depends on the choice of starting vectors $(r_0|$ and $|p_0)$ used to build it. This is reflected in the observation that such a space does not, in general, include the whole domain of the operator $\hat{L}$. More specifically, the operators $\hat{L}_0^{\mathrm{LZ}} = \hat{I}^{\mathrm{LZ}}\,\hat{L}\,\hat{I}^{\mathrm{LZ}}$ and $\hat{L}$ are not necessarily equivalent from a global point of view, i.e., in the space $\mathcal{G}^>$. Nonetheless, they are equivalent in the dual Lanczos vector space $\mathcal{G}^>_{\mathrm{LZ}}$.

In essence, a dual Lanczos transformation may be used to transform the retarded dynamics of a system into the closed retarded subdynamics of a dynamically invariant subspace. Of course, this dynamics is not necessarily equivalent to the retarded dynamics of the system. Nonetheless, knowledge of this subdynamics is sufficient for determining the properties of interest, provided we have properly biased the dual Lanczos vector space with the relevant dynamical information through our choice of $(r_0|$ and $|p_0)$.

Since the irrelevant information is discarded in a dual Lanczos transformation, it is not, in general, possible to recover the retarded system dynamics by performing an "inverse dual Lanczos transformation." The discarded information is irretrievable. This statment is embodied in the relation

$$\theta(t)\,\exp(-\hat{L}t) \neq \theta(t)\,\hat{X}^P\exp\big(-\hat{L}_0^{\mathrm{LZ}}t\big)\hat{X}^R. \tag{638}$$

The above relation applies unless $\hat{X}^P[\hat{X}^R]$ is the inverse of $\hat{X}^R[\hat{X}^P]$ with $\hat{X}^P\hat{X}^R = \hat{X}^R\hat{X}^P = \hat{I}$. If the latter set of relations hold, the inequality must be replaced by an equality. For such cases, the dual Lanczos vector space $\mathcal{G}^>_{\mathrm{LZ}}$ is defined in the whole domain of the operator $\hat{L}$ and $\mathcal{G}^>_{\mathrm{LZ}} = \mathcal{G}^>$.

5. Projected Transition Operators and Generalized Polynomials

The recursion relations given by Eqs. (625) and (626) may be simplified by introducing the transition operators

$$\hat{L}_{s+1}^{P,\mathrm{LZ}} = \hat{Q}_{s+1}^{\mathrm{LZ}}\hat{L} \tag{639}$$

and

$$\hat{L}_{s+1}^{R,\mathrm{LZ}} = \hat{L}\,\hat{Q}_{s+1}^{\mathrm{LZ}}, \tag{640}$$

where $\hat{Q}_{s+1}^{\mathrm{LZ}} = \hat{I}^{\mathrm{LZ}} - \hat{P}_{s+1}^{\mathrm{LZ}}$ is the complement of the projection operator

$$\hat{P}_{s+1}^{\mathrm{LZ}} = \sum_{j=0}^{s} |p_j)(r_j|. \tag{641}$$

Introducing the operators defined by Eqs. (639) and (640), we rewrite Eqs. (625) and (626) as

$$\beta_{s+1}|p_{s+1}) = \hat{L}_{s+1}^{P,\mathrm{LZ}}|p_s) \tag{642}$$

and

$$\beta_{s+1}(r_{s+1}| = (r_s|\hat{L}_{s+1}^{R,\mathrm{LZ}}. \tag{643}$$

In addition to the transition operators given by Eqs. (639) and (640), it is useful to work with the projected transition operator

$$\hat{L}_{s+1}^{\mathrm{LZ}} = \hat{Q}_{s+1}^{\mathrm{LZ}}\hat{L}\,\hat{Q}_{s+1}^{\mathrm{LZ}} \tag{644a}$$

$$= \hat{Q}_{s+1}^{\mathrm{LZ}}\hat{L}_{s+1}^{R,\mathrm{LZ}} \tag{644b}$$

$$= \hat{L}_{s+1}^{P,\mathrm{LZ}}\hat{Q}_{s+1}^{\mathrm{LZ}}. \tag{644c}$$

The operator $\hat{L}_{s+1}^{\mathrm{LZ}}$ is the projection of $\hat{L}$ on the dual Lanczos subspace orthogonal to the subspace spanned by the first $s+1$ dual Lanczos basis vectors. Since $\hat{L}_0^{\mathrm{LZ}} = \hat{I}^{\mathrm{LZ}}\hat{L}\,\hat{I}^{\mathrm{LZ}}$ by definition, $\hat{Q}_0^{\mathrm{LZ}} = \hat{I}^{\mathrm{LZ}}$ and $\hat{P}_0^{\mathrm{LZ}}$ is a null operator.

For some problems, one might find that a decent approximate treatment of the spectral and temporal properties of a system is obtained by working with the subspace spanned by the first $(s+1)$ dual Lanczos basis vectors. For such cases, we replace $\hat{L}_m^{\mathrm{LZ}}$ by its $(s+1)$-dimensional approximant $\hat{L}_m^{(s+1),\mathrm{LZ}}$.

The $(s+1)$-dimensional approximant $\hat{L}_m^{(s+1),\mathrm{LZ}}$ of the operator $\hat{L}_m^{\mathrm{LZ}}$ is defined by

$$\hat{L}_m^{(s+1),\mathrm{LZ}} = \hat{Q}_m^{(s+1),\mathrm{LZ}}\,\hat{L}\,\hat{Q}_m^{(s+1),\mathrm{LZ}}, \tag{645}$$

where $\hat{Q}_m^{(s+1),\mathrm{LZ}}$ is the orthogonal complement of $\hat{P}_m^{(s+1),\mathrm{LZ}} = \hat{P}_m^{\mathrm{LZ}}$ for $m = 0,\ldots,s+1$.

We can write $\hat{Q}_m^{(s+1),\mathrm{LZ}} = \hat{I}^{(s+1),\mathrm{LZ}} - \hat{P}_m^{(s+1),\mathrm{LZ}}$ for $m = 0,\ldots,s+1$, where $\hat{I}^{(s+1),\mathrm{LZ}} = \hat{P}_{s+1}^{\mathrm{LZ}}$ is the identity operator for the dual Lanczos subspace spanned by $\{(r_j|;\ j=0,\ldots,s\}$ and $\{|p_j);\ j=0,\ldots,s\}$. The operators $\hat{Q}_{s+1}^{(s+1),\mathrm{LZ}}$ and $\hat{P}_0^{(s+1),\mathrm{LZ}}$ are null operators by definition. Then $\hat{L}_{s+1}^{(s+1),\mathrm{LZ}}$ is a null operator and $\hat{Q}_0^{(s+1),\mathrm{LZ}} = \hat{I}^{(s+1),\mathrm{LZ}} = \hat{P}_{s+1}^{\mathrm{LZ}}$.

The operator $\hat{L}_0^{(s+1),\mathrm{LZ}}$ is simply the projection $\hat{P}_{s+1}^{\mathrm{LZ}}\hat{L}\,\hat{P}_{s+1}^{\mathrm{LZ}}$ of $\hat{L}$ on the dual Lanczos subspace spanned by the first $(s+1)$ dual Lanczos basis vectors. In other words, $\hat{L}_0^{(s+1),\mathrm{LZ}}$ is the $(s+1)$-dimensional approximant of $\hat{L}$ in the dual Lanczos representation.

It is convenient to introduce the function $\mathcal{A}_m(z) = |z\mathbf{Q}_m^{\mathrm{LZ}} + \mathbf{L}_m^{\mathrm{LZ}}|$, where z is a complex variable and $|z\mathbf{Q}_m^{\mathrm{LZ}} + \mathbf{L}_m^{\mathrm{LZ}}|$ is the determinant of the matrix $z\mathbf{Q}_m^{\mathrm{LZ}} + \mathbf{L}_m^{\mathrm{LZ}}$ formed by the operator $z\hat{Q}_m^{\mathrm{LZ}} + \hat{L}_m^{\mathrm{LZ}}$ in the dual Lanczos representation. $\mathcal{A}_m(z=-\lambda) = |-\lambda\mathbf{Q}_m{}^{\mathrm{LZ}} + \mathbf{L}_m^{\mathrm{LZ}}|$ is the characteristic function of the projected transition operator $\hat{L}_m^{\mathrm{LZ}}$ in the dual Lanczos subspace orthogonal to the subspace spanned by the first m dual Lanczos basis vectors.

The $(s+1)$-dimensional approximant $\mathcal{A}_m^{(s+1)}(z)$ of $\mathcal{A}_m(z)$ is defined by $\mathcal{A}_m^{(s+1)}(z) = |z\mathbf{Q}_m^{(s+1),\mathrm{LZ}} + \mathbf{L}_m^{(s+1),\mathrm{LZ}}|$, which represents a polynomial in z of degree $s+1-m$. Here, $\mathcal{A}_m^{(s+1)}(z=-\lambda) = |-\lambda\mathbf{Q}_m^{(s+1),\mathrm{LZ}} + \mathbf{L}_m^{(s+1),\mathrm{LZ}}|$ is the characteristic function of the projected transition operator $\hat{L}_m^{(s+1),\mathrm{LZ}}$ in the dual Lanczos subspace spanned by $\{(r_j|;\ j=m,\ldots,s\}$ and $\{|p_j);\ j=m,\ldots,s\}$. The quantity $\mathcal{A}_m^{(s+1)}(z)$ is equivalent to $\mathcal{A}_m(z)$ when $(s+1)$ is the actual dimensionality of the dual Lanczos vector space. We shall refer to the quantities $\{\mathcal{A}_m(z)\}$ and $\{\mathcal{A}_m^{(s+1)}(z)\}$ as generalized polynomials.

The generalized polynomials $\{\mathcal{A}_m^{(s+1)}(z)\}$ satisfy the recursion relations

$$\mathcal{A}_m^{(s+1)}(z) = (z+\alpha_m)\mathcal{A}_{m+1}^{(s+1)}(z) - \beta_{m+1}^2\mathcal{A}_{m+2}^{(s+1)}(z) \tag{646}$$

and

$$\mathcal{A}_m^{(s+1)}(z) = (z+\alpha_s)\mathcal{A}_m^{(s)}(z) - \beta_S^2\mathcal{A}_m^{(s-1)}(z), \tag{647}$$

where $\mathcal{A}_{s+1}^{(s+1)}(z) \equiv 1$, $\mathcal{A}_s^{(s+1)}(z) = (z+\alpha_s)$, and $\mathcal{A}_m^{(s+1)}(z) = 0$ for $m < 0$ and $m > s+1$.

The generalized polynomial $\mathcal{A}_m^{(s+1)}(z)$ can be written in the conventional form of a power series in z:

$$\mathcal{A}_m^{(s+1)}(z) = \sum_{j=0}^{s+1-m} d_m^{(s+1,j)} z^j, \tag{648}$$

where the expansion coefficients $\{d_m^{(s+1,j)}\}$ are given by $d_m^{(s+1,j)} = d_m^{(s,j-1)} + \alpha_s d_m^{(s,j)} - \beta_s^2 d_m^{(s-1,j)}$ for $j = 2=0,\ldots,s-1-m$; $d_m^{(s+1,s-m)} = d_m^{(s,s-1-m)} + \alpha_s$ for $j = s-m$; and $d_m^{(s+1,s+1-m)} = 1$ for $j = s+1-m$. Alternatively, the expansion coefficients are given by $d_m^{(s+1,j)} = d_m^{(s+1,j-1)} + \alpha_m d_{m+1}^{(s+1,j)} - \beta_{m+1}^2 d_{m+2}^{(s+1,j)}$ for $j = 0,\ldots,s-1-m$; $d_m^{(s+1,s-m)} = d_{m+1}^{(s+1,s-1-m)} + \alpha_m$ for $j = s-m$; and $d_m^{(s+1,s+1-m)} = 1$ for $j = s+1-m$.

The actual determination of the expansion coefficients in Eq. (648) requires the Lanczos parameters. These parameters may be determined by making use of Eqs. (625) and (626) to carry out a dual Lanczos transformation or by using the moments $\{M_l\}$, which are defined by $M_l = (-1)^l(r_0|\hat{L}^l|p_0)$. The Lanczos parameters are connected to the moments through the relations $\alpha_s = (\Pi_{m=0}^s \beta_m^2)^{-1}\sum_{j,j'=0}^s d_0^{(s,j)} d_0^{(s,j')} M_{j+j'+1}$ and $\beta_{s+1}^2 = -(\Pi_{m=0}^s \beta_m^2)^{-1}\sum_{j,j'=0}^{s+1} d_0^{(s+1,j)} d_0^{(s+1,j')} M_{j+j'}$, where $\beta_0 \equiv i$.

6. Extraction and Utilization of Dynamically Embedded Information

Independent of whether the underlying dynamics of a system is represented by a reversibleor irreversible classical

TABLE I Basic Quantities Appearing in Formal Expressions for Spectral and Temporal Properties of a System

Symbol	Formal expression	Identification
$G_{\psi,\chi}(t)$	$G^{>}_{\psi,\chi}(t)+G^{<}_{\psi,\chi}(t)$	Decomposition of time–dependent quantity $G_{\psi,\chi}(t)$ into retarded and advanced components
$G^{>}_{\psi,\chi}(t)$	$\theta(t)(\psi\vert\exp(-\hat{L}t)\vert\chi)$	Retarded component of $G_{\psi,\chi}(t)$
$G^{<}_{\psi,\chi}(t)$	$\theta(-t)(\psi\vert\exp(\hat{\tilde{L}}t)\vert\chi)$	Advanced component of $G_{\psi,\chi}(t)$
$g^{>}_{\psi,\chi}(z)$	$(\psi\vert(z\hat{I}+\hat{L})^{-1}\vert\chi)$	Laplace transform $\int_0^\infty dt\exp(-zt)G^{>}_{\psi,\chi}(t)$ of $G^{>}_{\psi,\chi}(t)$,where Rez > 0
$g^{F}_{\psi,\chi}(i\omega)$	$\lim_{\varepsilon\to 0^+}[g^{>}_{\psi,\chi}(i\omega+\varepsilon)+g^{>}_{\tilde{\psi},\tilde{\chi}}(-i\omega+\varepsilon)]$	Fourier transform $\int_{-\infty}^{+\infty} dt\exp(-i\omega t)G_{\psi,\chi}(t)$ of $G_{\psi,\chi}(t)$
$\vert\chi_t)$	$\vert\chi_t^{>})+\vert\chi_t^{<})$	Decomposition of dynamical vector $\vert\chi_t)$ into retarded and advanced components
$\vert\chi_t^{>})$	$\theta(t)\exp(-\hat{L}t)\vert\chi)$	Retarded component of $\vert\chi_t)$
$\vert\chi_t^{<})$	$\theta(-t)\exp(\hat{\tilde{L}}t)\vert\chi)$	Advanced component of $\vert\chi_t)$
$(\psi_t\vert$	$(\psi_t^{>}\vert+(\psi_t^{<}\vert$	Decomposition of dynamical vector $(\psi_t\vert$ into retarded and advanced components
$(\psi_t^{>}\vert$	$\theta(t)(\psi\vert\exp(-\hat{L}t)$	Retarded component of $(\psi_t\vert$
$(\psi_t^{<}\vert$	$\theta(-t)(\psi\vert\exp(\hat{\tilde{L}}t)$	Advanced component of $(\psi_t\vert$

or quantum dynamical model, we can usually express the spectral and temporal properties of a system in terms of the basic quantities displayed in Table I, where $(\psi\vert$ and $\vert\chi)$ are the pertinent dynamical vectors for the properties of interest. The transition operator $\hat{L}$ appearing in these basic quantities must be replaced by an appropriately projected form, such as $\hat{L}_{QQ}=\hat{Q}\hat{L}\hat{Q}$, when the underlying dynamics corresponds to a projected subdynamics of the system dynamics. Such a replacement is also required in all other relevant relations appearing in the formal apparatus of dual Lanczos transformation theory.

Within the context of dual Lanczos transformation theory, the basic steps involved in the determination of the spectral and temporal properties of interest are as follows: (i) Build a dual Lanczos vector space embedded with the appropriate dynamical information. (ii) Extract the information embedded in the dual Lanczos vector space. (iii) Utilize the extracted information to determine the spectral and temporal properties of interest.

In order to build a dual Lanczos vector space $\mathcal{G}^{>}_{\mathrm{LZ}}$, we focus our attention on some time-dependent quantity $A^{>}_{0,0}(t)$ and its Laplace transform $a^{>}_{0,0}(z)$:

$$A^{>}_{0,0}(t)=\theta(t)(r_0\vert\exp(-\hat{L}t)\vert p_0) \tag{649a}$$

$$=\theta(t)(r_0\vert\hat{I}^{\mathrm{LZ}}\exp(-\hat{L}t)\hat{I}^{\mathrm{LZ}}\vert p_0) \tag{649b}$$

$$=\theta(t)(r_0\vert\exp\bigl(-\hat{L}_0^{\mathrm{LZ}}t\bigr)\vert p_0) \tag{649c}$$

and

$$a^{>}_{0,0}=(r_0\vert(z\hat{I}+\hat{L})^{-1}\vert p_0) \tag{650a}$$

$$=(r_0\vert\hat{I}^{\mathrm{LZ}}(z\hat{I}+\hat{L})^{-1}\hat{I}^{\mathrm{LZ}}\vert p_0) \tag{650b}$$

$$=(r_0\vert\bigl(z\hat{I}^{\mathrm{LZ}}+\hat{L}_0^{\mathrm{LZ}}\bigr)^{-1}\vert p_0), \tag{650c}$$

where $(r_0\vert$ and $\vert p_0)$ are the starting vectors for building the dual Lanczos vector space $\mathcal{G}^{>}_{\mathrm{LZ}}$.

As indicated earlier, we require $(r_0\vert$ and $\vert p_0)$ to be of such a character that $(r_0\,\vert\,p_0)=1$ and $(r_0\vert\hat{L}^n\vert p_0)$ is well defined for all n. Provided these restrictions are satisfied, $A^{>}_{0,0}(t)$ may represent the retarded component of some normalized $[\lim_{t\to 0^+}A^{>}_{0,0}(t)=1]$ dynamical variable, distribution, ensemble average of a dynamical variable, autocorrelation function, cross-correlation function, or some linear combination of such quantities.

The choice of starting vectors $(r_0\vert$ and $\vert p_0)$ is usually biased in such a way that $A^{>}_{0,0}(t)$, $a^{>}_{0,0}(z)$, $(r^{>}_{0,t}\vert=\theta(t)(r_0\vert\exp(-\hat{L}t)$, or $\vert p^{>}_{0,t})=\theta(t)\exp(-\hat{L}t)\vert p_0)$ represent some quantity or linear combination of quantities that one is actually interested in. If this is not possible, $(r_0\vert$ and $\vert p_0)$ are selected in such a way that the dual Lanczos vector space built with these vectors bears the relevant dynamical information for determining the properties of interest. The dual Lanczos vector space $\mathcal{G}^{>}_{\mathrm{LZ}}$ bears sufficient dynamical information for determining all of the quantities displayed in Table I when $(\psi\vert$, $(\tilde{\psi}\vert$, $\vert\chi)$, and $\vert\tilde{\chi})$ lie in this space, i.e., $(\psi\,\vert\,\hat{I}^{\mathrm{LZ}}=(\psi\vert$, $(\tilde{\psi}\vert\hat{I}^{\mathrm{LZ}}=(\tilde{\psi}\vert$, $\vert\chi)=\hat{I}^{\mathrm{LZ}}\vert\chi)$, and $\hat{I}^{\mathrm{LZ}}\vert\tilde{\chi})=\vert\tilde{\chi})$.

Provided $(\psi\vert$ and $\vert\chi)$ lie in $\mathcal{G}^{>}_{\mathrm{LZ}}$, the time-reversed dynamical vectors $(\tilde{\psi}\vert$ and $\vert\tilde{\chi})$ will lie in $\mathcal{G}^{>}_{\mathrm{LZ}}$ when $(\psi\vert$ and $\vert\chi)$ possess definite time-reversal parity and/or $\hat{I}^{\mathrm{LZ}}$ is time-reversal invariant, i.e., $\hat{\tilde{I}}^{\mathrm{LZ}}=\hat{I}^{\mathrm{LZ}}$. If $(\tilde{\psi}\vert$ and $\vert\tilde{\chi})$ do not lie in $\mathcal{G}^{>}_{\mathrm{LZ}}$, it is necessary to build a second dual Lanczos vector space bearing the relevant dynamical information. For the sake of simplicity, we shall assume that a second dual Lanczos vector space is not required.

The extraction of the dynamical information embedded in the dual Lanczos vector space $\mathcal{G}^{>}_{\mathrm{LZ}}$ is accomplished by forming the projected operators

$$\theta(t)\hat{I}^{\mathrm{LZ}}\exp(-\hat{L}t)\hat{I}^{\mathrm{LZ}}=\sum_{j,k}\vert p_j)A^{>}_{j,k}(t)(r_k\vert \tag{651}$$

and

$$\hat{I}^{\mathrm{LZ}}(z\hat{I}+\hat{L})^{-1}\hat{I}^{\mathrm{LZ}}=\sum_{j,k}\vert p_j)a^{>}_{j,k}(z)(r_k\vert. \tag{652}$$

TABLE II Basic Quantities Determined via the Extraction and Utilization of Dynamically Embedded Information

Quantity[a]	Formal expression[b]
$G^{>}_{\psi,\chi}(t)$	$\sum_{j,k}(\psi\|p_j)A^{>}_{j,k}(t)(r_k\|\chi)$
$G^{<}_{\tilde\psi,\tilde\chi}(t)$	$\sum_{j,k}(\tilde\psi\|p_j)A^{>}_{j,k}(-t)(r_k\|\tilde\chi)$
$g^{>}_{\psi,\chi}(z)$	$\sum_{j,k}(\psi\|p_j)a^{>}_{j,k}(z)(r_k\|\chi)$
$g^{>}_{\tilde\psi,\tilde\chi}(z)$	$\sum_{j,k}(\tilde\psi\|p_j)a^{>}_{j,k}(z)(r_k\|\tilde\chi)$
$\|\chi_t^{>}\rangle$	$\sum_{j,k}\|p_j)A^{>}_{j,k}(t)(r_k\|\chi)$
$\|\chi_t^{<}\rangle$	$\sum_{j,k}\|\tilde p_j)A^{>}_{j,k}(-t)(\tilde r_k\|\chi)$
$(\psi_t^{>}\|$	$\sum_{j,k}(\psi\|p_j)A^{>}_{j,k}(t)(r_k\|$
$(\psi_t^{<}\|$	$\sum_{j,k}(\psi\|\tilde p_j)A^{>}_{j,k}(-t)(\tilde r_k\|$

[a] See Table I for the definition of the basic quantities.
[b] $(\tilde\psi\|p_j)=(\psi\|\tilde p_j)$ and $(r_k\|\tilde\chi)=(\tilde r_k\|\chi)$, where $\|\tilde p_j)$ and $(\tilde r_k\|$ are the time-reversed forms of $\|p_j)$ and $(r_k\|$, respectively.

The utilization of the extracted information to determine the spectral and temporal properties of interest is realized by making use of the information embedded in these operators. The results of applying this procedure to the quantities in Table I are given in Table II. The evaluation of the expressions in Table II is a much simpler task than the evaluation of the expressions in Table I by virtue of the fact that $\theta(t)\,\hat I^{\rm LZ}\exp(-\hat L t)\hat I^{\rm LZ} = \theta(t)\exp(-\hat L_0^{\rm LZ}t)$ and $\hat I^{\rm LZ}(z\hat I+\hat L)^{-1}\hat I^{\rm LZ}=(z\hat I^{\rm LZ}+\hat L_0^{\rm LZ})^{-1}$ represent the closed subdynamics of a dynamically invariant subspace bearing only the relevant dynamical information.

The time-dependent quantity $A^{>}_{j,k}(t)$ and the spectral function $a^{>}_{j,k}(z)$ appearing in Eqs. (651) and (652) and in the basic quantities displayed in Table II are given by

$$A^{>}_{j,k}(t) = \theta(t)\,(r_j|\exp(-\hat L t)\,|p_k) \tag{653a}$$

$$= \theta(t)(r_j|I^{\rm LZ}\exp(-\hat L t)\,I^{\rm LZ}|p_k) \tag{653b}$$

$$= \theta(t)\,(r_j|\exp\bigl(-\hat L_0^{\rm LZ}t\bigr)|p_k) \tag{653c}$$

and

$$a^{>}_{j,k}(z) = (r_j|(z\hat I+\hat L)^{-1}\,|p_k) \tag{654a}$$

$$= (r_j|I^{\rm LZ}(z\hat I+\hat L)^{-1}I^{\rm LZ}|p_k) \tag{654b}$$

$$= (r_j|\bigl(z\hat I^{\rm LZ}+\hat L_0^{\rm LZ}\bigr)^{-1}|p_k). \tag{654c}$$

$a^{>}_{j,k}(z)$ is the Laplace transform of $A^{>}_{j,k}(t)$ when $\mathrm{Re}\,z>0$. The quantities $A^{>}_{j,k}(t)$ and $a^{>}_{j,k}(z)$ satisfy the symmetry relations $A^{>}_{j,k}(t)=A^{>}_{k,j}(t)$ and $a^{>}_{j,k}(z)=a^{>}_{k,j}(z)$.

The quantity $A^{>}_{j,k}(t)$ can be determined by employing either the inverse Laplace transformation

$$A^{>}_{j,k}(t) = (2\pi i)^{-1}\int_{C^{>}} dz\,\exp(zt)\,a^{>}_{j,k}(z) \tag{655}$$

or the transformation

$$A^{>}_{j,k}(t) = (2\pi i)^{-1}\int_{C} dz\,\exp(zt)\,a_{j,k}(z) \tag{656}$$

based on an operator analogue of the Cauchy integral formula.

The contour $C^{>}$ in Eq. (655) is a straight line $z=\gamma$ parallel to the imaginary axis, where the real number γ is chosen so that $z=\gamma+i\omega$ lies to the right of the singularities of $a^{>}_{j,k}(z)$ but is otherwise arbitrary. The contour C in Eq. (656) encloses the singularities of the spectral function

$$a_{j,k}(z) = \left[(-1)^{(j+k+1)}\left(\prod_{m=0}^{j}\beta_m\right)\left(\prod_{n=0}^{k}\beta_n\right)\right]^{-1} \times \mathcal{A}_0^{(j)}(z)\mathcal{A}_0^{(k)}(z)a_{0,0}(z). \tag{657}$$

The spectral function $a_{0,0}(z)$ is identical in form to $a^{>}_{0,0}(z)$ [see Eqs. (654a)–(654c)] except that $a_{0,0}(z)$ is by construction defined in the entire complex plane rather than only in the right half. Note that $a_{j,k}(z)$ satisfies the symmetry relation $a_{j,k}(z)=a_{k,j}(z)$.

Although the contour $C^{>}$ in inverse Laplace transformations is usually replaced by the Bromwich contour, which does enclose the singularities of $a^{>}_{j,k}(z)$, the spectral functions $a^{>}_{j,k}(z)$ and $a_{j,k}(z)$ are generally not equivalent. [$a_{0,0}(z)=a^{>}_{0,0}(z)$ is an exception when $\mathrm{Re}\,z>0$]. The Laplace transform $a^{>}_{j,k}(z)$ is defined only in the right half of the complex plane, while the spectral function $a_{j,k}(z)$ is by construction defined throughout the complex plane. In fact, the spectral functions $a^{>}_{j,k}(z)$ and $a_{j,k}(z)$ are connected by the relation

$$a^{>}_{j,k}(z) = (2\pi i)^{-1}\int_{C^*} dz' a_{j,k}(z')/(z-z'), \tag{658}$$

where C^* is a closed contour that encloses the singularities of $a_{j,k}(z)$ and excludes the point z. Equation (658) may be used to construct the Laplace transform $a^{>}_{j,k}(z)$ from the spectral function $a_{j,k}(z)$ by performing a suitable contour integration.

An alternative and simpler way to determine the Laplace transform $a^{>}_{j,k}(z)$ is to employ the relations

$$a^{>}_{j,k}(z) = \left[(-1)^{(j+k+1)}\left(\prod_{m=0}^{j}\beta_m\right)\left(\prod_{n=0}^{k}\beta_n\right)\right]^{-1} \times \bigl\{\mathcal{A}_0^{(j)}(z)\mathcal{A}_0^{(k)}(z)\bigl[a^{>}_{0,0}(z)-\mathcal{A}_1^{(k)}(z)/\mathcal{A}_0^{(k)}(z)\bigr]\bigr\} \tag{659}$$

for $k\geq j$ and $k\geq 1$ and

$$a^{>}_{j,k}(z) = \left[(-1)^{(j+k+1)}\left(\prod_{m=0}^{j}\beta_m\right)\left(\prod_{n=0}^{k}\beta_n\right)\right]^{-1} \times \bigl\{\mathcal{A}_0^{(j)}(z)\mathcal{A}_0^{(k)}(z)\bigl[a^{>}_{0,0}(z)-\mathcal{A}_1^{(j)}(z)/\mathcal{A}_0^{(j)}(z)\bigr]\bigr\} \tag{660}$$

for $j \geq k$ and $j \geq 1$. Comparing Eq. (657) with Eqs. (659) and (660) for Re $z > 0$, one can see that $a^{>}_{j,k}(z)$ and $a_{j,k}(z)$ differ by an analytic function.

Other useful forms for the Laplace transform $a^{>}_{j,k}(z)$ are given by

$$a^{>}_{j,k}(z) = \left[(-1)^{(j+k)}\left(\prod_{m=j}^{k}\beta_m\right)\Big/\beta_j\right] \times \left[\mathcal{A}_0^{(j)}(z)\mathcal{A}_{k+1}(z)\right]/\mathcal{A}_0(z) \tag{661}$$

for $k \geq j$ and

$$a^{>}_{j,k}(z) = \left[(-1)^{(j+k)}\left(\prod_{m=k}^{j}\beta_m\right)\Big/\beta_k\right] \times \left[\mathcal{A}_0^{(k)}(z)\mathcal{A}_{j+1}(z)\right]/\mathcal{A}_0(z) \tag{662}$$

for $k \leq j$.

As indicated in Tables I and II, one can determine the Fourier transform $g^{\mathrm{F}}_{\psi,\chi}(i\omega)$ of the time-dependent quantity $G_{\psi,\chi}(t)$ by making use of our knowledge of the spectral functions $\{a^{>}_{j,k}(z)\}$. More specifically, we can determine $g^{\mathrm{F}}_{\psi,\chi}(i\omega)$ by employing

$$g^{\mathrm{F}}_{\psi,\chi}(i\omega) = \lim_{\epsilon\to 0^+}\sum_{j,k}\left[(\psi|p_j)a^{>}_{j,k}(i\omega+\epsilon)(r_k|\chi) + (\tilde{\psi}|p_j)a^{>}_{j,k}(-i\omega+\epsilon)(r_k|\tilde{\chi})\right]. \tag{663}$$

Provided $(\psi|$, $(\tilde{\psi}|$, $|\chi)$, and $|\tilde{\chi})$ lie in the dual Lanczos vector space at hand, the result given by Eq. (663) applies regardless of the nature of the underlying dynamics. Nonetheless, a dramatically simpler form may be used for the case of reversible systems, i.e., when the symmetry relation $\hat{\tilde{L}} = -\hat{L}$ holds. For the case of reversible systems, we can write $g^{\mathrm{F}}_{\psi,\chi}(i\omega)$ as follows:

$$g^{\mathrm{F}}_{\psi,\chi}(i\omega) = \sum_{j,k}(\psi|p_j)a^{\mathrm{F}}_{j,k}(i\omega)(r_k|\chi), \tag{664}$$

where the spectral function

$$a^{\mathrm{F}}_{j,k}(i\omega) = \left[(-1)^{(j+k+1)}\left(\prod_{m=0}^{j}\beta_m\right)\left(\prod_{n=0}^{k}\beta_n\right)\right]^{-1} \times \mathcal{A}_0^{(j)}(i\omega)\mathcal{A}_0^{(k)}(i\omega)a^{\mathrm{F}}_{0,0}(i\omega) \tag{665}$$

is the Fourier transform $a^{\mathrm{F}}_{j,k}(i\omega) = \int_{-\infty}^{+\infty} dt\,\exp(-i\omega t)\,A_{j,k}(t)$ of $A_{j,k}(t) = (r_j|\exp(-\hat{L}t)|p_k)$.

The determination of $a^{\mathrm{F}}_{j,k}(i\omega)$ by means of Eq. (665) requires $a^{\mathrm{F}}_{0,0}(i\omega)$. The latter spectral function may be determined by employing Eq. (663) provided the time-reversed dynamical vectors $(\tilde{r}_0|$ and $|\tilde{p}_0)$ lie in the dual Lanczos vector space at hand. As indicated earlier, this will be true when $(r_0|$ and $|p_0)$ possess definite time-reversal parity. Assuming this to be the case, the normalization $(r_0\,|\,p_0) = 1$ requires $(r_0|$ and $|p_0)$ to be of the same parity. It follows that we can determine $a^{\mathrm{F}}_{0,0}(i\omega)$ by employing the relation

$$a^{\mathrm{F}}_{0,0}(i\omega) = \lim_{\epsilon\to 0^+}\left[a^{>}_{0,0}(i\omega+\epsilon) + a^{>}_{0,0}(-i\omega+\epsilon)\right] \tag{666}$$

regardless of the nature of the underlying dynamics when $(r_0|$ and $|p_0)$ possess definite and equal time-reversal parity.

Taking advantage of the fact that the spectral function $a^{>}_{0,0}(z)$ may be written as either a continued fraction

$$a^{>}_{0,0}(z) = \frac{1}{z+\alpha_0-}\cdots\frac{-\beta_s^2}{z+\alpha_s-}\cdots \tag{667}$$

or the ratio

$$a^{>}_{0,0}(z) = \mathcal{A}_1(z)/\mathcal{A}_0(z) \tag{668}$$

of generalized polynomials, we can write the Fourier transform $a^{\mathrm{F}}_{0,0}(i\omega)$ as

$$a^{\mathrm{F}}_{0,0}(i\omega) = \lim_{\epsilon\to 0^+}\left\{\left[\frac{1}{(i\omega+\epsilon)+\alpha_0-}\cdots\frac{-\beta_s^2}{(i\omega+\epsilon)+\alpha_s-}\cdots\right] + \left[\frac{1}{(-i\omega+\epsilon)+\alpha_0-}\cdots\frac{-\beta_s^2}{(-i\omega+\epsilon)+\alpha_s-}\cdots\right]\right\} \tag{669}$$

or

$$a^{\mathrm{F}}_{0,0}(i\omega) = \lim_{\epsilon\to 0^+}\{[\mathcal{A}_1(i\omega+\epsilon)/\mathcal{A}_0(i\omega+\epsilon)] + [\mathcal{A}_1(-i\omega+\epsilon)/\mathcal{A}_0(-i\omega+\epsilon)]\}. \tag{670}$$

With the results given thus far, one can proceed to determine the spectral and temporal properties of a system by utilizing the universal formulas displayed in Table II. As indicated earlier, this approach does not require the solution of the system equations of motion. Moreover, one does not have to use matrix diagonalization techniques, which, in general, do not apply.

For the case in which $\hat{L}_0^{\mathrm{LZ}}$ possesses distinct eigenvalues or, equivalently, for the case in which $\mathcal{A}_0(z)$ possesses distinct zeros, we can write

$$A^{>}_{j,k}(t) = \theta(t)\sum_{l_0}D_{j,k,l_0}\exp\left(-\lambda_{0,l_0}t\right) \tag{671}$$

and

$$a^{>}_{j,k}(z) = \sum_{l_0}D_{j,k,l_0}\left(z+\lambda_{0,l_0}\right)^{-1}, \tag{672}$$

where

$$D_{j,k,l_0} = \left[(-1)^{(j+k+1)}\left(\prod_{m=0}^{j}\beta_m\right)\left(\prod_{n=0}^{k}\beta_n\right)\right]^{-1} \times \mathcal{A}_0^{(j)}\left(-\lambda_{0,l_0}\right)\mathcal{A}_0^{(k)}\left(-\lambda_{0,l_0}\right)C_{0,l_0}^2, \quad (673)$$

with

$$C_{0,l_0}^2 = \mathcal{A}_1\left(-\lambda_{0,l_0}\right)\Big/\left[d\mathcal{A}_0\left(-\lambda_{0,l_0}\right)/dz\right]. \quad (674)$$

The index l_0 runs over the eigenvalues $\{\lambda_{0,l_0}\}$ of $\hat{L}_0^{\mathrm{LZ}}$, which are connected to the zeros $\{z_{0,l_0}\}$ of $\mathcal{A}_0(z)$ by the relation $\lambda_{0,l_0} = -z_{0,l_0}$. Here, C_{0,l_0}^2 is the square of the matrix element C_{0,l_0} of the complex orthogonal transformation matrix that would be obtained in the diagonalization of $\mathbf{L}_0^{\mathrm{LZ}}$ by means of a complex orthogonal transformation. For the case of finite-dimensional spaces, the $\{C_{0,l_0}\}$ and $\{\lambda_{0,l_0}\}$ required in the evaluation of Eqs. (673) and (674) can be obtained by a QR transformation scheme involving the rotation of a vector with the components δ_{0,l_0} for $l_0 = 0, 1, \ldots$. In general, both C_{0,l_0}^2 and λ_{0,l_0} are complex.

As indicated earlier, it might not be possible and/or desirable to obtain and work with the full dual Lanczos vector space generated with the starting vectors $(r_0|$ and $|p_0)$. For some problems, one might find that a decent approximate treatment of the spectral and temporal properties of a system is obtained by working only with the subspace spanned by the first $(s+1)$ dual Lanczos basis vectors. Such an approximate treatment is obtained by setting $\beta_j = 0$ for $j \geq s+1$ and replacing $\hat{L}_m^{\mathrm{LZ}}$ and $\mathcal{A}_m(z)$ by their $(s+1)$-dimensional approximants $\hat{L}_m^{(s+1),\mathrm{LZ}}$ and $\mathcal{A}_m^{(s+1)}(z)$ in all of the relevant expressions.

7. Memory Function Hierarchies, Continued Fractions, and Padé Approximants

The time evolution of $A_{0,0}^{>}(t)$ [see Eqs. (649a)–(649c)] can be described in terms of a chain of non-Markovian equations of motion

$$\frac{d}{dt}K_m^{>}(t) = -\alpha_m K_m^{>}(t) - \int_0^t dt' K_{m+1}^{>}(t-t')K_m^{>}(t') \quad (675)$$

for the members $\{K_m^{>}(t); m = 0, 1, \ldots\}$ of the memory function hierarchy of $A_{0,0}^{>}(t)$, where the memory function $K_m^{>}(t)$ is given by

$$K_m^{>}(t) = -\left(r_{m-1}\middle|\hat{L}_m^{R,\mathrm{LZ}}\exp\left(-\hat{L}_m^{\mathrm{LZ}}t\right)\hat{L}_m^{P,\mathrm{LZ}}\middle|p_{m-1}\right) \quad (676a)$$

$$= -\beta_m^2\left(r_m\middle|\exp\left(-\hat{L}_m^{\mathrm{LZ}}t\right)\middle|p_m\right), \quad (676b)$$

with $t > 0$. Equation (676b) assumes the form of $A_{0,0}^{>}(t)$ when $m = 0$, i.e., $A_{0,0}^{>}(t) = K_0^{>}(t)$. (Recallthat $\beta_0^2 = -1$.)

The time evolution of the mth member $K_m^{>}(t)$ of the memory function hierarchy may be determined by using the relation

$$K_m^{>}(t) = (2\pi i)^{-1}\int_{C^{>}} dz\, \exp(zt)\,\mathcal{K}_m^{>}(z), \quad (677)$$

where

$$\mathcal{K}_m^{>}(z) = -\beta_m^2[\mathcal{A}_{m+1}(z)/\mathcal{A}_m(z)] \quad (678)$$

is the Laplace transform of $K_m^{>}(t)$ and the contour $C^{>}$ is a straight line $z = \gamma$ parallel to the imaginary axis, with the real number γ chosen so that $z = \gamma + i\omega$ lies to the right of the singularities of $\mathcal{K}_m^{>}(z)$ but is otherwise arbitrary.

For the case in which $\hat{L}_m^{\mathrm{LZ}}$ possesses distinct eigen values or, equivalently, for the case in which $\mathcal{A}_m(z)$ possesses distinct zeros, we can write

$$K_m^{>}(t) = -\theta(t)\beta_m^2\sum_{k_m} C_{m,k_m}^2\exp\left(-\lambda_{m,k_m}t\right) \quad (679)$$

and

$$\mathcal{K}_m^{>}(z) = -\beta_m^2\sum_{k_m} C_{m,k_m}^2\left(z+\lambda_{m,k_m}\right)^{-1}, \quad (680)$$

where

$$C_{m,k_m}^2 = \mathcal{A}_{m+1}\left(-\lambda_{m,k_m}\right)\Big/\left[d\mathcal{A}_m\left(-\lambda_{m,k_m}\right)/dz\right]. \quad (681)$$

The index k_m runs over the eigenvalues $\{\lambda_{m,k_m}\}$ of $\hat{L}_m^{\mathrm{LZ}}$, which are connected to the zeros $\{z_{m,k_m}\}$ of $\mathcal{A}_m(z)$ by the relation $\lambda_{m,k_m} = -z_{m,k_m}$. Here C_{m,k_m}^2 is the square of the matrix element C_{m,k_m} of the complex orthogonal transformation matrix that would be obtained in the diagonalization of $\mathbf{L}_m^{\mathrm{LZ}}$ by means of a complex orthogonal transformation. For the case of finite dimensional spaces, the $\{C_{m,k_m}\}$ and $\{\lambda_{m,k_m}\}$ required in the evaluation of Eqs. (680) and (681) can be obtained by a QR transformation scheme involving the rotation of a vector with the components δ_{m,k_m} for $k_m = m, m+1, \ldots$.

The Laplace transform of Eq. (675) provides a continued fraction representation of $\mathcal{K}_m^{>}(z)$:

$$\mathcal{K}_m^{>}(z) = \frac{-\beta_m^2}{\left[(z+\alpha_m) + \mathcal{K}_{m+1}^{>}(z)\right]}. \quad (682)$$

The above result is equivalent to the continued fraction representation of $a_{0,0}^{>}(z)$ given by Eq. (667) when $m = 0$.

It follows from Eqs. (678) and (682) that the $(s+1)$-dimensional approximant $\mathcal{K}_m^{>,(s+1)}(z)$ of $\mathcal{K}_m^{>}(z)$ is given by

$$\mathcal{K}_m^{>,(s+1)}(z) = -\beta_m^2\left[\mathcal{A}_{m+1}^{(s+1)}(z)\Big/\mathcal{A}_m^{(s+1)}(z)\right] \quad (683)$$

or

$$\mathcal{K}_m^{>,(s+1)}(z) = \frac{-\beta_m^2}{\left[(z+\alpha_m) + \mathcal{K}_{m+1}^{>,(s+1)}(z)\right]}, \quad (684)$$

where $\mathcal{K}_m^{>,(s+1)}(z)=0$ for $m \geq s+1$ and $\mathcal{K}_0^{>,(s+1)}(z)=a_{0,0}^{>,(s+1)}(z)$. The form given by Eq. (683) represents the $[s-m/s+1-m]$ Padé approximant of $\mathcal{K}_m^{>}(z)$.

8. Generalized Langevin Equations

The time evolution of the dynamical vectors $(r_{0,t}^{>}|=\theta(t)\,(r_0|\exp(-\hat{L}t)$ and $|p_{0,t}^{>})=\theta(t)\exp(-\hat{L}t)\,|\,p_0)$ may be described in terms of chains of generalized Langevin equations given by

$$\frac{d}{dt}\big(r_{m,t}^{>}\big| = -\alpha_m\big(r_{m,t}^{>}\big| - \int_0^t dt' K_{m+1}^{>}(t-t')\big(r_{m,t'}^{>}\big| - \beta_{m+1}\big(r_{m+1,t}^{>}\big| \tag{685}$$

and

$$\frac{d}{dt}\big|p_{m,t}^{>}\big) = -\alpha_m\big|p_{m,t}^{>}\big) - \int_0^t dt'\, K_{m+1}^{>}(t-t')\big|p_{m,t'}^{>}\big) - \beta_{m+1}\big|p_{m+1,t}^{>}\big) \tag{686}$$

for $t>0$, where $(r_{m,t}^{>}|=\theta(t)\,(r_m|\exp(-\hat{L}_m^{\mathrm{LZ}}t)$ and $|p_{m,t}^{>})=\theta(t)\,\exp(-\hat{L}_m^{\mathrm{LZ}}t)|p_m)$. Interpreting $(r_{m+1,t}^{>}|$ and $|p_{m+1,t}^{>})$ as spontaneous noise, we see that non-Markovian retardation in the generalized Langevin equations is due to correlated noise, which is described by the memory function $K_{m+1}^{>}(t-t')$.

The solutions to the chains of generalized Langevin equations can be written in the forms

$$\big(r_{0,t}^{>}\big| = \sum_j A_{0,j}^{>}(t)\,(r_j| \tag{687}$$

and

$$\big|p_{0,t}^{>}\big) = \sum_j |p_j) A_{j,0}^{>}(t). \tag{688}$$

For the case in which $A_{0,0}^{>}(t)$ represents some equilibrium time correlation function, the $\{A_{0,j}^{>}(t)=A_{j,0}^{>}(t)\}$ correspond to a set of linearly independent equilibrium time correlation functions characterizing the spontaneous equilibrium fluctuations that are coupled to and drive the time evolution of $(r_{0,t}^{>}|$ and $|p_{0,t}^{>})$.

9. Spectral Domains, Time Scales, and Memory Effects

Understanding the character of the spectral and temporal properties of a dynamical system requires information about the distribution of the eigenvalues of the projected transition operators $\{\hat{L}_m^{\mathrm{LZ}}\}$ and their $(s+1)$-dimensional approximants $\{\hat{L}_m^{(s+1),\mathrm{LZ}}\}$. Often the real part of an eigenvalue is assumed to describe the rate of a physically well-defined process.

The author and coworkers have introduced a number of theorems from the analytic theory of polynomial equations and perturbation theory for the purpose of gaining insight into the distribution of eigenvalues by simply knowing the Lanczos parameters. These theorems, which include the Gershgorin circle theorems, enable one to construct spectral domains in the complex plane to which the eigenvalues of $\{\hat{L}_m^{\mathrm{LZ}}\}$ and $\{\hat{L}_m^{(s+1),\mathrm{LZ}}\}$ are confined. It has been shown that an analysis of the size and configuration of the spectral domains not only provides bounds on the eigenvalues, but also provides considerable insight into spectral and temporal character, enabling one to make rigorous statements about separations in time scales and memory effects.

10. Illustrative Application

As a simple illustrative application of dual Lanczos transformation theory, let us consider the problem of determining the Fourier transform $\sigma(\vec{k},\omega)$ of the incoherent scattering function $S(\vec{k},t)$ associated with a particle in a spatially homogeneous, isotropic environment and executing motion described by Fokker–Planck dynamics.

We write

$$\sigma(\vec{k},\omega) = \int_{-\infty}^{+\infty} dt \exp(-i\omega t) S(\vec{k},t), \tag{689}$$

where

$$S(\vec{k},t) = S^{>}(\vec{k},t) + S^{<}(\vec{k},t), \tag{690}$$

with

$$S^{<}(\vec{k},t) = \theta(t)\,(\exp(+i\vec{k}\cdot\vec{q})|\exp(-\hat{L}t)| \times \exp(-i\vec{k}\cdot\vec{q})\,\rho_{\mathrm{eq}}) \tag{691}$$

and

$$S^{<}(\vec{k},t) = \theta(-t)\,(\exp(+i\vec{k}\cdot\vec{q})|\exp(+\hat{\tilde{L}}t)| \times \exp(-i\vec{k}\cdot\vec{q})\,\rho_{\mathrm{eq}}). \tag{692}$$

In the above, $\hat{L}$ is an abstract operator corresponding to the Fokker–Planck transition operator $L(\vec{p},\vec{q})$ for a damped particle executing motion in a spatially homogeneous, isotropic environment, where $\vec{p}$ and $\vec{q}$, respectively, are the momentum and coordinate vectors for the particle of interest. The vectors $(\exp(+i\vec{k}\cdot\vec{q})|$ and $|\exp(-i\vec{k}\cdot\vec{q})\rho_{\mathrm{eq}})$ are dynamical vectors corresponding to the phase functions $\exp(+i\vec{k}\cdot\vec{q})$ and $\exp(-i\vec{k}\cdot\vec{q})\rho_{\mathrm{eq}}(\vec{p},\vec{q})$, where $\rho_{\mathrm{eq}}(\vec{p},\vec{q})$ is the equilibrium probability density for the particle of interest and $\vec{k}$ denotes the change in the wavevector of the incident photon/neutron in a light/neutron scattering experiment.

Choosing $(r_0|=(\exp(+i\vec{k}\cdot\vec{q})|$ and $|p_0)=|\exp(-i\vec{k}\cdot\vec{q})\rho_{\mathrm{eq}})$ as starting vectors, we can build a dual Lanczos vector space $\mathcal{G}_{\mathrm{LZ}}^{>}$ embedded with sufficient information to determine the quantities $\sigma(\vec{k},\omega)$, $S(\vec{k},t)$, $S^{>}(\vec{k},t)$, and

$S^<(\vec{k}, t)$. Making this choice of starting vectors, we find that $\sigma(\vec{k}, \omega)$ and $S^>(\vec{k}, t)$ assume the forms of $a^{\mathrm{F}}_{0,0}(i\omega)$ and $A^>_{0,0}(t)$, respectively.

Implementing the procedure of extraction and utilization of dynamically embedded information on a computer using symbolic manipulation techniques, we obtain the following exact results:

$$S^>(\vec{k}, t) = \theta(t)\, \exp\left\{\left(\frac{k_{\mathrm{B}}T|\vec{k}|^2}{m\xi^2}\right)[1 - \xi t - \exp(-\xi t)]\right\}, \tag{693}$$

$$S^<(\vec{k}, t) = \theta(-t) \exp\left\{\left(\frac{k_{\mathrm{B}}T|\vec{k}|^2}{m\xi^2}\right)[1 + \xi t - \exp(\xi t)]\right\}, \tag{694}$$

and

$$\sigma(\vec{k}, \omega) = 2\exp\left(\frac{k_{\mathrm{B}}T|\vec{k}|^2}{m\xi^2}\right) \sum_{l=1}^{\infty} \left\{ \frac{(-1)^l}{l!} \left[\frac{k_{\mathrm{B}}T|\vec{k}|^2}{m\xi^2}\right]^l \times \left[\frac{\left(k_{\mathrm{B}}T|\vec{k}|^2/m\xi\right) + l\xi}{\omega^2 + \left[\left(k_{\mathrm{B}}T|\vec{k}|^2/m\xi\right) + l\xi\right]^2}\right]\right\}. \tag{695}$$

SEE ALSO THE FOLLOWING ARTICLES

CHEMICAL THERMODYNAMICS • INFORMATION THEORY • MECHANICS, CLASSICAL • QUANTUM MECHANICS • STATISTICS, FOUNDATIONS • THERMODYNAMICS

BIBLIOGRAPHY

Berne, B. J. (ed.). (1977). "Statistical Mechanics, Part B: Time Dependent Processes," Plenum, New York.

Callen, H. B. (1962). "Thermodynamics," Wiley, New York.

de Boer, J., and Uhlenbeck, G. E. (eds.). (1962). "Studies in Statistical Mechanics," Vol. 1, North-Holland, Amsterdam.

de Groot, S. R., and Mazur, P. (1984). "Non-equilibrium Thermodynamics," Dover, New York.

Forster, D. (1975). "Hydrodynamic Fluctuations, Broken Symmetry, and Correlation Functions," Benjamin, Reading, MA.

Glansdorff, P., and Prigogine, I. (1971). "Thermodynamics of Structure, Stability and Fluctuations," Wiley, New York.

Grabert, H. (1982). "Projection Operator Techniques in Nonequilibrium Statistical Mechanics," Springer-Verlag, Berlin.

Jancel, R. (1963). "Foundations of Classical and Quantum Statistical Mechanics," Pergamon Press, London.

Katz, A. (1967). "Principles of Statistical Mechanics, The Information Theory Appproach," Freeman, San Francisco.

Kondepudi, D., and Prigogine, I. (1980). "Modern Thermodynamics, From Heat Engines to Dissipative Structures," Wiley, New York.

Kubo, R., Toda, M., and Hashitsume, H. (1985). "Statistical Physics II, Nonequilibrium Statistical Mechanics," Springer-Verlag, Berlin.

Louisell, W. H. (1973). "Quantum Statistical Properties of Radiation," Wiley, New York.

McQuarrie, D. A. (1976). "Statistical Mechanics," Harper and Row, New York.

Prigogine, I. (1980). "From Being to Becoming, Time and Complexity in the Physical Sciences," Freeman, San Francisco.

Reichl, L. E. (1980). "A Modern Course in Statistical Physics," University of Texas Press, Austin, TX.

Toda, M., Kubo, R., and Saito, N. (1972). "Statistical Physics I, Equilibrium Statistical Mechanics," Springer-Verlag, Berlin.

Wax, N. (ed.). (1954). "Selected Papers on Noise and Stochastic Processes," Dover, New York.

Yourgrau, W., van der Merwe, A., and Raw, G. (1966). "Treatise on Irreversible and Statistical Thermophysics, An Introduction to Nonclassical Thermodynamics," Macmillan, New York.

Zubarev, D. N. (1974). "Nonequilibrium Statistical Thermodynamics," Consultants Bureau, New York.

Steam Tables

Allan H. Harvey
National Institute of Standards and Technology

GLOSSARY

Critical point For a pure substance, the upper limit of the vapor–liquid saturation curve where the equilibrium vapor and liquid phases become identical and the compressibility of the fluid becomes infinite. For water, the critical point occurs at a temperature of approximately 374°C and a pressure of approximately 22 MPa.

Formulation A mathematical equation or set of equations from which a desired quantity or set of quantities (such as the thermodynamic properties of water) can be calculated.

Saturation A condition where two phases of a substance (in most common usage, the vapor and the liquid) are in thermodynamic equilibrium. Some thermodynamic variables, such as the temperature and pressure, have identical values in the two phases at saturation; other properties, such as density and enthalpy, have different values in each phase.

Skeleton tables Accepted values of properties presented at specific values (usually round numbers) of temperature and pressure. These are based on analysis and interpolation of data at nearby conditions, and an estimate of the uncertainty of each value at each point is usually included.

STEAM TABLES is the traditional name for tabulations of the thermodynamic properties of water in both its vapor (steam) and liquid states. This information is of particular importance for the steam power-generation industry, but many other industrial processes make use of water in some form and therefore need reliable values of its properties. In addition, water is widely used in research, and some of these uses require highly accurate representations of its properties. Modern "steam tables" are for the most part no longer printed tables, but are mathematical formulations implemented in computer programs.

I. INTRODUCTION

The need for standardized representation of properties of water and steam is most apparent in the power industry, where electricity is generated by passing large amounts of steam through turbines. The thermodynamic properties of

water and steam are vital to the design of equipment and to the evaluation of its performance. Small differences in the properties used will produce small changes in calculated quantities such as thermal efficiencies; however, because of the magnitude of the steam flows, these differences can translate into millions of dollars. It is therefore essential that all parties in the steam power industry, particularly bidders and purchasers in contracts, use a uniform set of properties in order to prevent any party from having an unfair advantage.

For other industries, such as petroleum refining and chemical manufacturing, standardization of water properties is less important, but there is still a need for reliable thermodynamic values for process design, optimization, and operation. In some cases, the complexity of the formulation is an issue, because the properties must be evaluated within iterative calculations. Thus, there is an incentive to keep the formulations as simple as possible without sacrificing significant accuracy or consistency.

Scientific and engineering research also requires highly accurate values for properties of water. This is not only for work where water is used directly, but also because of the widespread use of water as a calibration fluid.

In this article, we review the historical development of steam tables and then describe the current international standards as maintained by the International Association for the Properties of Water and Steam (IAPWS). While the primary focus of steam tables (and therefore of this article) is thermodynamic properties (density, enthalpy, entropy, etc.), other properties (such as viscosity, thermal conductivity, and dielectric constant) are also of some importance and will be mentioned briefly.

II. EARLY HISTORY OF STEAM POWER AND STEAM TABLES

About 2000 years ago, a figure showing a workable steam reaction turbine was included in a book by Hero of Alexandria. He showed other ways in which steam, or other hot gases, could be used to do mechanical work, and used the boiler, valve, and piston (basic components of a steam engine) at various places in his book. In spite of this promising start, the use of steam for power never advanced significantly in antiquity; it remained for later generations to develop its potential.

Although early engines of Papin, Savery, and Newcomen paved the way, it was James Watt in the late 1700s who made steam power an industrial success. Watt greatly improved the design of steam engines and took advantage of the improved metal-working techniques then available. Early in his career, Watt measured the temperature and pressure of saturated steam and constructed a curve through the points to permit interpolation. In a sense this curve was the first steam table; however, the science of Watt's day was inadequate to make much use of such data.

In the 1840s, V. Regnault (with some assistance from a young William Thomson, who later became Lord Kelvin) produced a set of careful measurements of the properties of steam. These data and others from Regnault's laboratory provided a foundation for the development and application of the new science of thermodynamics by Thomson, Clausius, and Rankine. By the late 19th century, steam tables based on Regnault's data began to appear, and in 1900 Callendar devised a thermodynamically consistent set of equations for treating steam data. Further steam tables soon appeared based on Callendar's equations; Mollier published the first steam tables in modern form.

The proliferation of steam tables soon became a problem, because different tables used different data. The differences between tables were particularly serious at high temperatures and pressures, where industrial interest was concentrated in the quest for increased thermodynamic efficiency. In addition to uncertainties in design, the different tables made it difficult to compare designs from different manufacturers. It became clear that international standardization was necessary in order to put all parties in the industry on a fair, consistent, and physically sound basis.

III. INTERNATIONAL STANDARDIZATION

The first conference directed at producing international agreement on steam tables was held in London in 1929. After a second (Berlin, 1930) and a third (New York, 1934) conference, agreement was reached on a set of "skeleton tables." These tables consisted of a rectangular grid of temperatures and pressures, with values of specific volume and enthalpy determined at each point by interpolation of the surrounding data. Values were similarly determined along the vapor–liquid saturation curve. An estimated uncertainty was assigned to each value at each point. These tables and their supporting data were the basis for the widely used steam tables book of J. H. Keenan and F. G. Keyes, published in 1936, which became the *de facto* standard for engineering calculations for many years.

After World War II, international standardization efforts resumed. Improvements in power-generation technology required reliable properties at pressures and temperatures beyond the range covered by Keenan and Keyes, and new data (notably from laboratories in the Soviet Union) became available. At the Sixth International Conference on the Properties of Steam (New York, 1963), a new set of skeleton tables was approved, covering an expanded range of pressures and temperatures.

By 1963, it was recognized that the growing use of computers for engineering calculations made it desirable

to represent the properties of water and steam by equations in addition to skeleton tables. An International Formulation Committee (IFC) was appointed to develop a consistent set of equations that reproduced the 1963 skeleton table values within their tolerances. The main product of this effort was "The 1967 IFC Formulation for Industrial Use" (known as IFC-67). This formulation, and the printed steam tables based on it, replaced the Keenan and Keyes tables as the standard for industrial calculations for the next 30 years. Because of its association with the book of the same name, it is sometimes called the 1967 "ASME Steam Tables" formulation, although the American Society of Mechanical Engineers was only one of several participants in the international effort, and steam tables based on IFC-67 were published in several countries in addition to the United States.

At about the same time, the need was recognized for a permanent organization to manage the international conferences and the maintenance and improvement of property standards. In 1968, the International Association for the Properties of Steam (IAPS) was established; in 1989 the name was changed to the International Association for the Properties of Water and Steam (IAPWS). IAPWS, which now has 11 member countries, continues to work to improve international standards for water and steam properties for use in science and industry. It meets annually, and sponsors an International Conference on the Properties of Water and Steam every five years. Some of the activities of IAPWS are discussed in the following sections.

IV. STANDARDS FOR GENERAL AND SCIENTIFIC USE

There are actually two different audiences for steam tables. The steam power industry, which historically provided the impetus for standardized steam tables, needs a formulation that can be calculated relatively quickly by computer for use in iterative design calculations. It also needs a standard that is fixed for many years, because switching from one property formulation to another involves much adjustment of software that uses steam properties and impacts areas such as contracting and testing of equipment where large sums of money are at stake. Once the accuracy attains a level sufficient for most engineering purposes, the industry is willing to forego additional accuracy for the sake of speed and stability.

However, there are other users in industry for whom speed and stability are not significant issues. In addition, researchers need to have the most accurate properties available for their work. IAPWS therefore has two separate tracks of standards. Formulations "for industrial use," such as IFC-67, are designed for computational speed and are intended to remain the standard for use in the power industry for decades. In contrast, formulations "for general and scientific use" are intended to be kept at the state of the art, giving the best possible representation of the best experimental data in conjunction with theoretical constraints regardless of how complicated they must be or how frequently they are updated.

The first thermodynamic property formulation for general and scientific use was adopted in 1968, but was never widely used. In 1984, IAPS replaced it with a formulation developed by L. Haar, J. S. Gallagher, and G. S. Kell. The Haar–Gallagher–Kell (HGK) formulation, codified in the *NBS/NRC Steam Tables*, saw widespread use as a standard for scientific work and for some engineering applications. It used a single thermodynamic function (the Helmholtz energy as a function of temperature and density) to cover a wide range of temperatures and pressures. This guarantees thermodynamic consistency and prevents the discontinuities inherent in formulations that use different equations for different pressure/temperature regions. While the HGK formulation was not the first to take this approach, it was the first such approach to be adopted as an international standard.

As better data became available and small flaws were found in the HGK formulation, IAPWS in the early 1990s began an organized effort to produce a replacement. A task group, led by W. Wagner, evaluated the available experimental data and worked on producing an improved formulation; others tested the formulation exhaustively. The final result was the "IAPWS Formulation 1995 for the Thermodynamic Properties of Ordinary Water Substance for General and Scientific Use," which we shall refer to as IAPWS-95.

The structure of IAPWS-95 is a single equation for the Helmholtz energy as a function of temperature and density. All thermodynamic properties can be obtained in a consistent manner from differentiation and manipulation of that equation. The equation was also forced to meet certain theoretical constraints such as a correct approach to the ideal-gas limit at low densities, and it closely approximates the correct behavior as water's critical point is approached.

IAPWS-95 is now the state of the art and the international standard for representing water's thermodynamic properties at temperatures from its freezing point to 1000°C and at pressures up to 1000 MPa. It also extrapolates in a physically meaningful manner outside this range, including the supercooled liquid water region. The uncertainties in the properties produced by IAPWS-95 are comparable to those of the best available experimental data; this is quite accurate in some cases (for example, relative uncertainty of 10^{-6} for liquid densities at atmospheric pressure and near-ambient temperatures) and less so where the data are less certain (for example, relative uncertainty of 2×10^{-3} for most vapor heat

capacities). Values of properties for saturation states and the property change upon vaporization, as generated by IAPWS-95, are shown in a typical steam tables format in Table I.

Formulations for general and scientific use have also been adopted for other properties of water. The most industrially important of these are probably the viscosity and the thermal conductivity, although some other properties such as the static dielectric constant and the refractive index are important in research. There are also IAPWS formulations for some properties of heavy water (deuterium oxide, D_2O).

Table II shows all the properties for which IAPWS has thus far adopted official formulations (known as "releases"). Copies of specific IAPWS releases may be obtained at no charge from www.iapws.org or by requesting them from the Executive Secretary of IAPWS. Currently, the Executive Secretary is:

Dr. R. B. Dooley
Electric Power Research Institute
3412 Hillview Avenue
Palo Alto, California 94304

V. STANDARDS FOR INDUSTRIAL USE

As mentioned earlier, the steam power industry requires a standard formulation that is both stable (in the sense of not changing for tens of years) and computationally fast. For 30 years, the IFC-67 formulation mentioned in Section III fulfilled that need. Through the years, however, some deficiencies in IFC-67 became apparent. Probably the worst problems related to inconsistencies at the boundaries between the regions of pressure–temperature space in which different equations defined the formulation. These inconsistencies can cause problems in iterative calculations near the boundaries. Also, with improvements in optimization methods and computer technology, it was believed that the computational speed of IFC-67 could be surpassed. In addition, there was a desire (driven by the increasing use of combustion turbines) to add standard properties for steam at temperatures higher than the upper limit of 800°C for IFC-67.

Therefore, in parallel to (and slightly behind) the development of the IAPWS-95 standard for general and scientific use, IAPWS undertook an effort to develop a new formulation for industrial use that would replace IFC-67. This effort, led by a development task group chaired by W. Wagner and a testing task group chaired by K. Miyagawa, resulted in the adoption of a new standard in 1997 called "IAPWS Industrial Formulation 1997 for the Thermodynamic Properties of Water and Steam" (abbreviated IAPWS-IF97).

The structure of IAPWS-IF97 is shown in Fig. 1. It consists of five regions defined in terms of pressure and temperature. The heavy solid line (Region 4) is the vapor–liquid saturation curve, represented by a single equation giving the saturation pressure as a function of temperature (and vice versa). The compressed liquid (Region 1) and the superheated vapor (Region 2) are represented by equations giving the Gibbs energy as a function of pressure and temperature (the most convenient independent variables for typical power-industry calculations). Other thermodynamic functions are obtained by appropriate differentiation of the Gibbs energy function. A Gibbs energy equation is also used in Region 5, which covers the high-temperature range needed for combustion turbines. In Region 3, which includes the area around the critical point, a Helmholtz energy function is used with density and temperature as independent variables (because pressure and temperature do not work well as independent variables near the critical point). Careful efforts were made to ensure that the values of the thermodynamic properties at either side of the region boundaries matched within tight tolerances.

Figure 1 also indicates the so-called "backward" equations in Regions 1 and 2, which allow the temperature to be obtained directly from pressure and enthalpy, or pressure and entropy, without iteration. The backward equations were made to reproduce the results from iterative solution of the "forward" equations (which have p and T as independent variables) within close tolerances. The backward equations in IAPWS-IF97 provide a great increase in speed for calculations (common in the power industry) where pressure is known in combination with either entropy or enthalpy. The backward equations necessarily introduce some inconsistency compared to exact solution of the forward equations, but this is negligible for most purposes. If greater consistency is desired at the expense of speed, the backward equations can be used as initial guesses for iterative solution of the forward equations to the required precision.

The accuracy of IAPWS-IF97 is for the most part only slightly less than that of the IAPWS-95 formulation for general and scientific use. In fact, rather than being fitted to experimental data, IAPWS-IF97 was fitted to the IAPWS-95 formulation and therefore agrees with it closely.

For industrial users, switching from IFC-67 to IAPWS-IF97 can be a major effort. Especially in the design and testing of large power-generation equipment, the relatively small changes in properties can produce numbers sufficiently different to have a large economic impact. Other aspects of the design and testing process, including software, which have been "tuned" to give the right results for IFC-67 properties, must therefore be readjusted to be

TABLE I Thermodynamic Properties of Water in Saturated Liquid and Vapor States as Calculated from the IAPWS-95 Formulation

t (°C)	Pressure MPa	Volume, cm^3/g			Enthalpy, kJ/kg			Entropy, kJ/(kg·K)			t (°C)
		v_L	Δv	v_V	h_L	Δh	h_V	s_L	Δs	s_V	
0.01	0.000 612	1.0002	205 990	205 991	0.00	2500.9	2500.9	0.0000	9.1555	9.1555	0.01
5	0.000 873	1.0001	147 010	147 011	21.02	2489.0	2510.1	0.0763	8.9486	9.0248	5
10	0.001 228	1.0003	106 302	106 303	42.02	2477.2	2519.2	0.1511	8.7487	8.8998	10
15	0.001 706	1.0009	77 874	77 875	62.98	2465.4	2528.3	0.2245	8.5558	8.7803	15
20	0.002 339	1.0018	57 756	57 757	83.91	2453.5	2537.4	0.2965	8.3695	8.6660	20
25	0.003 170	1.0030	43 336	43 337	104.83	2441.7	2546.5	0.3672	8.1894	8.5566	25
30	0.004 247	1.0044	32 877	32 878	125.73	2429.8	2555.5	0.4368	8.0152	8.4520	30
35	0.005 629	1.0060	25 204	25 205	146.63	2417.9	2564.5	0.5051	7.8466	8.3517	35
40	0.007 385	1.0079	19 514	19 515	167.53	2406.0	2573.5	0.5724	7.6831	8.2555	40
45	0.009 595	1.0099	15 251	15 252	188.43	2394.0	2582.4	0.6386	7.5247	8.1633	45
50	0.012 352	1.0121	12 026	12 027	209.3	2381.9	2591.3	0.7038	7.3710	8.0748	50
60	0.019 946	1.0171	7666.2	7667.2	251.2	2357.7	2608.8	0.8313	7.0769	7.9081	60
70	0.031 201	1.0228	5038.5	5039.5	293.1	2333.0	2626.1	0.9551	6.7989	7.7540	70
80	0.047 414	1.0291	3404.1	3405.2	335.0	2308.0	2643.0	1.0756	6.5355	7.6111	80
90	0.070 182	1.0360	2358.0	2359.1	377.0	2282.5	2659.5	1.1929	6.2853	7.4781	90
100	0.101 42	1.0435	1670.7	1671.8	419.2	2256.4	2675.6	1.3072	6.0469	7.3541	100
110	0.143 38	1.0516	1208.2	1209.3	461.4	2229.6	2691.1	1.4188	5.8193	7.2381	110
120	0.198 67	1.0603	890.15	891.21	503.8	2202.1	2705.9	1.5279	5.6012	7.1291	120
130	0.270 28	1.0697	666.93	668.00	546.4	2173.7	2720.1	1.6346	5.3918	7.0264	130
140	0.361 54	1.0798	507.37	508.45	589.2	2144.3	2733.4	1.7392	5.1901	6.9293	140
150	0.476 16	1.0905	391.36	392.45	632.2	2113.7	2745.9	1.8418	4.9953	6.8371	150
160	0.618 23	1.1020	305.68	306.78	675.5	2082.0	2757.4	1.9426	4.8066	6.7491	160
170	0.792 19	1.1143	241.48	242.59	719.1	2048.8	2767.9	2.0417	4.6233	6.6650	170
180	1.0028	1.1274	192.71	193.84	763.1	2014.2	2777.2	2.1392	4.4448	6.5840	180
190	1.2552	1.1415	155.22	156.36	807.4	1977.9	2785.3	2.2355	4.2704	6.5059	190
200	1.5549	1.1565	126.05	127.21	852.3	1939.7	2792.0	2.3305	4.0996	6.4302	200
210	1.9077	1.1727	103.12	104.29	897.6	1899.6	2797.3	2.4245	3.9318	6.3563	210
220	2.3196	1.1902	84.902	86.092	943.6	1857.4	2800.9	2.5177	3.7663	6.2840	220
230	2.7971	1.2090	70.294	71.503	990.2	1812.7	2802.9	2.6101	3.6027	6.2128	230
240	3.3469	1.2295	58.476	59.705	1037.6	1765.4	2803.0	2.7020	3.4403	6.1423	240
250	3.9762	1.2517	48.831	50.083	1085.8	1715.2	2800.9	2.7935	3.2785	6.0721	250
260	4.6923	1.2761	40.897	42.173	1135.0	1661.6	2796.6	2.8849	3.1167	6.0016	260
270	5.5030	1.3030	34.318	35.621	1185.3	1604.4	2789.7	2.9765	2.9539	5.9304	270
280	6.4166	1.3328	28.820	30.153	1236.9	1543.0	2779.9	3.0685	2.7894	5.8579	280
290	7.4418	1.3663	24.189	25.555	1290.0	1476.7	2766.7	3.1612	2.6222	5.7834	290
300	8.5879	1.4042	20.256	21.660	1345.0	1404.6	2749.6	3.2552	2.4507	5.7059	300
310	9.8651	1.4479	16.887	18.335	1402.2	1325.7	2727.9	3.3510	2.2734	5.6244	310
320	11.284	1.4990	13.972	15.471	1462.2	1238.4	2700.6	3.4494	2.0878	5.5372	320
330	12.858	1.5606	11.418	12.979	1525.9	1140.2	2666.0	3.5518	1.8903	5.4422	330
340	14.601	1.6376	9.143	10.781	1594.5	1027.3	2621.8	3.6601	1.6755	5.3356	340
350	16.529	1.7400	7.062	8.802	1670.9	892.7	2563.6	3.7784	1.4326	5.2110	350
360	18.666	1.8954	5.054	6.949	1761.7	719.8	2481.5	3.9167	1.1369	5.0536	360
370	21.044	2.215	2.739	4.954	1890.7	443.8	2334.5	4.1112	0.6901	4.8012	370
t_c[a]	22.064	3.106	0	3.106	2084.3	0	2084.3	4.4070	0	4.4070	t_c

[a] $t_c = 373.946$°C.

TABLE II IAPWS Releases for Calculating Properties of Water and Heavy Water[a]

Property	Date of latest version
Thermal conductivity[b]	1998
Viscosity[b]	1997
Refractive Index	1997
Static dielectric constant	1997
Thermodynamic properties (industrial use)	1997
Thermodynamic properties (general and scientific use)	1996
Surface tension	1994
Surface tension (D_2O)	1994
Melting and sublimation pressures	1993
Critical point properties	1992
Thermodynamic properties (D_2O)	1984
Viscosity and thermal conductivity (D_2O)	1984
Ion product	1980

[a] Copies of IAPWS Releases may be obtained by writing to the IAPWS Executive Secretary: Dr. R. B. Dooley, Electric Power Research Institute, 3412 Hillview Ave., Palo Alto, CA 94304.

[b] These releases contain formulations both for industrial use and for general and scientific use.

consistent with IAPWS-IF97. IAPWS, upon adopting the formulation, recommended a waiting period (which expired at the beginning of 1999) in which IAPWS-IF97 should not be used for contractual specifications, in order to allow users time to adjust. Similar adjustments and therefore a similar waiting period will likely be required whenever a successor to IAPWS-IF97 is adopted; however, the intention is for IAPWS-IF97 to remain the standard in the power industry for at least 20 years.

Property formulations for industrial use have also been generated for the viscosity and the thermal conductivity, as mentioned in Table II.

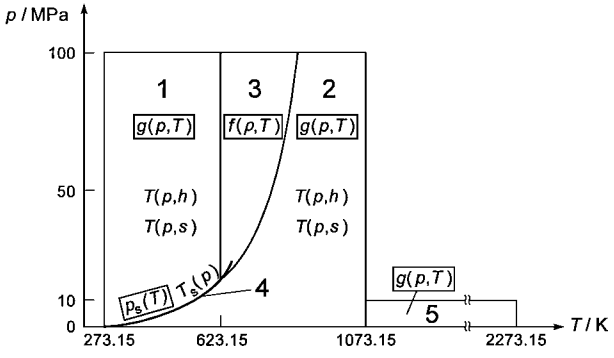


FIGURE 1 Regions of the IAPWS-IF97 standard for thermodynamic properties of water and steam for industrial use.

A final question to be addressed is when it is appropriate to use the industrial formulation (IAPWS-IF97), as opposed to the formulation for general and scientific use (IAPWS-95). In general, since IAPWS-95 is the state of the art, it (or any formulation for general and scientific use that might replace it in the future) should be used in all cases except those where IAPWS-IF97 is specifically required or preferable. IAPWS-IF97 is preferred in two cases:

- In the steam power industry, the industrial formulation (now IAPWS-IF97) is the industry standard for contracting and testing purposes. It therefore makes sense to use IAPWS-IF97 for calculations in all facets of the power industry.
- In any application where computing time is at a premium (and where the calculation of water properties consumes most of that time), IAPWS-IF97 may be preferred because of its much faster computing speed. An example would be finite-element calculations of steam flow in a turbine. In some cases, even IAPWS-IF97 might be too slow; an alternative in such cases (especially if the calculations are confined to a narrow range of conditions) is to generate a table of properties in advance and then use a table interpolation algorithm in the computations.

VI. FUTURE DIRECTIONS

Probably the most notable current direction for steam tables is the migration from printed tables to computer software. Most engineering design and research now uses computer-generated properties. Printed tables and charts formerly had to be detailed enough to permit interpolation with an accuracy suitable for design; now they are mostly relegated to the role of an auxiliary to be used for quick estimates when the computer is not handy, and therefore can be much less detailed. With the increased use of computers, it has become increasingly important to present water property standards in user-friendly software, either for standalone calculations or as something to be plugged into a spreadsheet or other application. There is also demand for implementations of steam property formulations in computer languages other than the traditional FORTRAN and for access to properties via the World Wide Web. Figure 2 shows a window from some modern "steam tables" software.

As new data are obtained and new theoretical understanding is gained, work will continue to improve the existing formulations for the properties of water and steam. Much of this work is organized by IAPWS. Current areas of focus include improvement of the formulations

1: L/V sat. T=300.0 to 500.0 [K]

	1 Temperature [K]	2 Pressure [MPa]	3 Density (L) [kg/m^3]	4 Density (V) [kg/m^3]	5 Enthalpy (L) [kJ/kg]	6 Enthalpy (V) [kJ/kg]
1	300.0	0.003537	996.5	0.02559	112.6	2550
2	320.0	0.01055	989.4	0.07166	196.2	2586
3	340.0	0.02719	979.5	0.1744	279.9	2621
4	360.0	0.06219	967.4	0.3786	363.8	2654
5	380.0	0.1289	953.3	0.7483	448.1	2686
6	400.0	0.2458	937.5	1.369	533.0	2716
7	420.0	0.4373	919.9	2.352	618.6	2742
8	440.0	0.7337	900.6	3.833	705.3	2765
9	460.0	1.171	879.6	5.983	793.4	2783
10	480.0	1.790	856.5	9.014	883.3	2796
11	500.0	2.639	831.3	13.20	975.4	2802

FIGURE 2 Sample screen from a modern "steam tables" database implementing the IAPWS-95 formulation (From A. H. Harvey, A. P. Peskin, and S. A. Klein, "NIST/ASME Steam Properties," NIST Standard Reference Database 10, Standard Reference Data Office, NIST, Gaithersburg, MD 20899; information also available at srdata@nist.gov or http://www.nist.gov/srd/nist10.htm).

for viscosity and thermal conductivity, and representation of water's thermodynamic behavior in accordance with theoretical constraints in the vicinity of its critical point.

IAPWS is also increasing its activities in application areas where water properties play a role. In the physical chemistry of aqueous solutions, efforts are devoted not only to those properties of pure water (such as the dielectric constant and ionization constant) that are important for solution chemistry, but also to the description of key properties of aqueous mixtures. Areas of interest include the partitioning of solutes between liquid water and steam and the properties of water/ammonia mixtures. In power-plant chemistry, IAPWS seeks to help industrial users apply fundamental knowledge and identifies key areas requiring further research. Documents called IAPWS Certified Research Needs (ICRN's) describe these needs; these documents are intended to serve as evidence to those who set research priorities that work in an area would be of significant use to industry. More information on current ICRN's may be obtained from the IAPWS Executive Secretary (see Section IV) or on the IAPWS Website (see following).

New data, new scientific capabilities, and new industrial needs will continue to shape the international steam properties community (and specifically IAPWS) as it seeks to maintain its core mission of developing steam tables and other water property standards while expanding into related areas to meet the needs of the power industry and of others who require accurate knowledge of the properties of water and aqueous mixtures. Up-to-date information on the activities of IAPWS may be found on its Website at www.iapws.org.

SEE ALSO THE FOLLOWING ARTICLES

CRITICAL DATA IN PHYSICS AND CHEMISTRY • THERMODYNAMICS • THERMOMETRY • WATER CONDITIONING, INDUSTRIAL

BIBLIOGRAPHY

Haar, L., Gallagher, J. S., and Kell, G. S. (1984). "NBS/NRC Steam Tables," Hemisphere Publishing Corporation, New York.

Harvey, A. H., and Parry, W. T. (1999). Keep Your "Steam Tables" Up to Date. *Chemical Eng. Progr.* **95**(11), 45.

Parry, W. T., Bellows, J. C., Gallagher, J. S., and Harvey, A. H. (2000). "ASME International Steam Tables for Industrial Use," ASME Press, New York.

Tremaine, P. R., Hill, P. G., Irish, D. E., and Balakrishnan, P. V. (eds.) (2000). "Steam, Water, and Hydrothermal Systems: Physics and Chemistry Meeting the Needs of Industry, Proceedings of the 13th International Conference on the Properties of Water and Steam," NRC Research Press, Ottawa.

Wagner, W., and Kruse, A. (1998). "Properties of Water and Steam," Springer-Verlag, Berlin.

White, Jr., H. J., Sengers, J. V., Neumann, D. B., and Bellows, J. C. (eds.) (1995). "Physical Chemistry of Aqueous Systems: Meeting the Needs of Industry, Proceedings of the 12th International Conference on the Properties of Water and Steam," Begell House, New York.

Thermoluminescence Dating

Geoff Duller
University of Wales, Aberystwyth

GLOSSARY

Annual dose The dose of ionizing radiation received by a sample per year during burial.

a-value The ratio between the luminescence signal induced by a given dose of alpha particles and that induced by the same dose of beta or gamma radiation.

Equivalent dose (ED) Dose of ionizing radiation received by a sample subsequent to the event being dated, as evaluated from measurements of natural and artificial luminescence. ED can also be denoted by the term D_E. An alternative term is palaeodose (P).

Fading Loss of luminescence during storage. Anomalous fading refers to loss in excess of the thermal fading predicted from measurement of the relevant electron trap depth and frequency factor.

Glow curve Plot of light intensity versus temperature as a sample is heated up during measurement of thermoluminescence (TL) (see Fig. 1a).

Natural luminescence The TL or OSL signal measured from a sample that is due to its exposure to ionizing radiation during burial. This is as opposed to "artificial luminescence" which is generated by exposure to radiation in the laboratory.

Optical decay curve Plot of light intensity versus time as a sample is stimulated by exposure to light during measurement of optically stimulated luminescence (OSL) (see Fig. 1b).

Palaeodose (P) See *Equivalent dose.*

Plateau test A plot of the ED calculated for a sample using different temperatures from the TL glow curve.

Preheating Annealing a sample at a fixed temperature for a fixed period of time prior to TL or OSL measurement. This has the effect of removing trapped charge from thermally unstable traps.

Residual signal The luminescence signal remaining after a zeroing event. A major advantage of OSL compared with TL is that for a given light exposure the residual signal is smaller for OSL measurements.

Saturation While exposure to radiation causes charge to accumulate in traps, the luminescence signal subsequently generated increases. However, beyond a certain point, additional radiation exposure does not cause more charge to be trapped since the trapping sites are full. At this point the sample is said to be saturated.

Zeroing Complete or partial elimination of previously acquired luminescence. It is the zeroing event that is

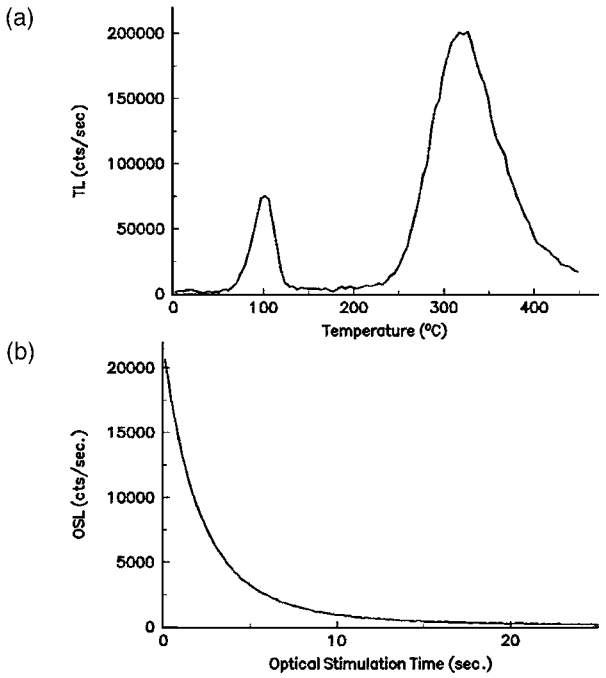


FIGURE 1 (a) A thermoluminescence (TL) glow curve showing two distinct peaks. (b) An optical decay curve generated during an OSL measurement.

normally dated by luminescence methods (see *Residual signal*).

FOLLOWING EXPOSURE to ionizing radiation (for example, alpha or beta particles, gamma radiation, or X-rays), some minerals (for example, quartz and feldspar), emit light when stimulated at some later time. The intensity of this *luminescence* is proportional to the dosage (amount) of ionizing radiation that has been absorbed since the last event that eliminated any previously acquired luminescence. For minerals in baked clay this event is firing (for example, of pottery), while for geological sediments the event is exposure to sunlight during transport and deposition. The ionizing radiation is provided by radioactive elements (e.g., ^{40}K, Th, and U) in the sample and its surroundings during burial, and cosmic rays. The age is obtained as:

$$\text{Age} = \text{ED/D} \tag{1}$$

where ED is the laboratory estimate of the radiation dose to which the sample has been exposed since the event being dated. This is known as the equivalent dose. D is the dose per year to the sample during burial (the annual dose). Luminescence measurements are used to determine ED, the equivalent dose, while emission counting or chemical measurements are used to calculate the annual dose.

I. PHYSICAL MECHANISM

A. The Basic Model

The details of the mechanism by which luminescence is produced in any given material are not well understood, and in general it is only for crystals grown in the laboratory with strict control of impurities that these details can be elucidated. However, the main features of luminescence dating can be discussed in terms of a simple model. In this, the free electrons produced by ionizing radiation are stored in traps where they remain in a metastable condition until the sample is stimulated, release being caused by increased lattice vibrations. Some of the released electrons find their way to luminescence centers, and incidental to the process of combining into such centers, photons are emitted. There are electron traps in the majority of natural ionic crystals and covalent solids, formed by defects in the lattice structure. The luminescence centers are associated with impurity atoms, and these determine the color of the luminescence emitted.

Electron traps are essentially localized regions where there is a deficit of negative charge or an excess of positive charge. The luminescence process can also result from holes being charge carriers, but in discussion it is convenient to talk only of electrons.

B. Stability

The length of time for which an electron remains trapped is one control on the age range over which luminescence dating can be applied. This period of time is determined by the energy, E, that is necessary to free the electron from the type of trap concerned and the frequency factor, s. The probability of escape per unit time is equal to the reciprocal of the lifetime, τ, where

$$\tau = \text{s}^{-1} \exp(E/kT) \tag{2}$$

where k is the Boltzmann's constant and T is the absolute temperature (degrees Kelvin). For the 325°C peak in quartz, $E = 1.69$ eV and $s = 10^{14}$ sec^{-1}, giving $\tau = 10^8$ yr at an ambient temperature of 15°C; lifetimes for other peaks occurring at temperatures above 300°C for this and other minerals are comparable.

Some minerals (for example feldspar) from some sources exhibit anomalous fading; namely, the observed lifetime is less than predicted by Eq. (2). The effect is because of an escape route not via the conduction band but directly to nearby centers. Two mechanisms have been proposed: (1) a localized transition via an excited state common to both the trap and center, and (2) wave-mechanical tunneling from the trap to the center. Reliable dating is not possible if (2) is dominant.

C. Thermal and Optical Stimulation

Measurement of a luminescence signal in the laboratory requires stimulation of the sample in order to evict charge from the metastable traps into the conduction band. This stimulation can be achieved in a number of ways, but the two used for dating are either by heating the sample to generate thermoluminescence (TL) or by exposure to light to generate optically stimulated luminescence (OSL). It is not always clear whether the charge evicted during OSL measurements can be directly linked to a specific part of that evicted during TL measurements.

For many dating applications the measurement of OSL instead of TL has a significant advantage. In dating geological materials, the event that is dated is the last exposure of the mineral grains to daylight during transport and deposition; this exposure will have reduced the luminescence signal to a low level. The effect of light upon the TL signal is variable, with some TL peaks being reduced more rapidly than others, but in almost all cases the TL signal is not reduced to zero and a considerable residual signal remains. Estimation of the residual signal at deposition then introduces an additional source of uncertainty. Through exposure to light, OSL signals tend to be reduced at least an order of magnitude more rapidly than those measured with TL, and the residual level after prolonged exposure to light is much smaller. The combined effect is that using measurements of OSL instead of TL permits dating of younger events, where the magnitude of the residual level becomes more significant, and also the dating of geological materials such as fluvial sands where the exposure to daylight at deposition may have been limited in duration.

II. AGE EVALUATION

A. Measurement of Luminescence

For dating, a luminescence signal can be obtained from both quartz and feldspars. Where grains between 100 and 300 μm are used, the quartz can be separated from the feldspars using chemical and physical processing. Alternatively, much finer grains (4 to 11 μm) may be analyzed, but in this case no attempt is made to separate different mineral types. These grain size ranges are selected because of the alpha dosimetry (see Section II.B).

Although the luminescence emitted by some bright geological minerals can be seen with the naked eye, the luminescence intensity encountered in samples to be dated is very faint. Not only is it necessary to use a highly sensitive low-noise photomultiplier, but it is also vital to discriminate against unwanted light. When measuring TL the primary concern is suppressing the incandescence generated by the sample and hotplate at high temperatures. This can be efficiently done by the use of broadband glass filters that exclude wavelengths in the red and infrared parts of the spectrum.

For OSL measurements the situation is rather different. Optical stimulation is commonly carried out either using infrared (~880 nm) or visible wavelengths (~450–520 nm). An intense OSL signal is observed from feldspars when stimulated in either waveband, while quartz only gives a bright signal when stimulated in the visible. The optical stimulation source used is commonly emitting 10^{10} times more photons than the sample being measured. Thus, it is vital to place glass filters in front of the photomultiplier that will reject the wavelengths used for stimulation, but not those emitted by the sample.

For measurement, the mineral grains of a sample are carried on a metal disk, usually 10 mm in diameter. Automated equipment incorporating a sample changer, radioactive source, hotplate, optical stimulation sources, and a photomultiplier is frequently used.

Since many of the luminescence signals that are observed are light sensitive, all sample preparation and measurement are carried out in subdued red light. This is a wavelength that has a negligible effect upon the signals normally observed.

B. Equivalent Dose Determination

Luminescence measurements are used to estimate the radiation dose to which the mineral grains have been exposed since the event being dated—the equivalent dose (ED), or palaeodose (P) in Eq. (1). Since there is no intrinsic relationship between radiation dose and the luminescence signal subsequently measured, each sample requires individual calibration. Two basic approaches have been used to determine the ED: additive dose and regenerative dose (Fig. 2).

The additive dose procedure involves separate measurements of the luminescence signal resulting from the radiation received by the sample during burial (its natural luminescence), and of the sum of that radiation dose and known doses administered in the laboratory. A mathematical function is then fitted to these data to characterize the growth of the luminescence signal with radiation exposure. This curve is known as a growth curve (Fig. 2a). The radiation dose during burial can then be calculated by extrapolating the growth curve to the residual luminescence signal remaining after exposure to daylight.

The regenerative dose procedure also starts with measurement of the natural luminescence signal, resulting from the radiation received by the sample during burial. The sample is then exposed to daylight in order to remove the luminescence signal, and known laboratory doses are administered that regenerate the luminescence signal. A

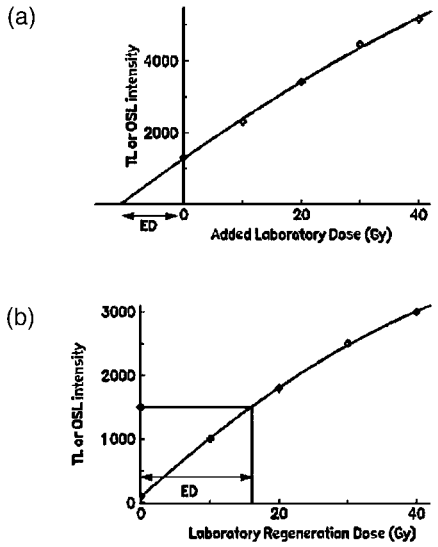


FIGURE 2 Additive dose (a) and regenerative dose (b) procedures for determination of the equivalent dose (ED).

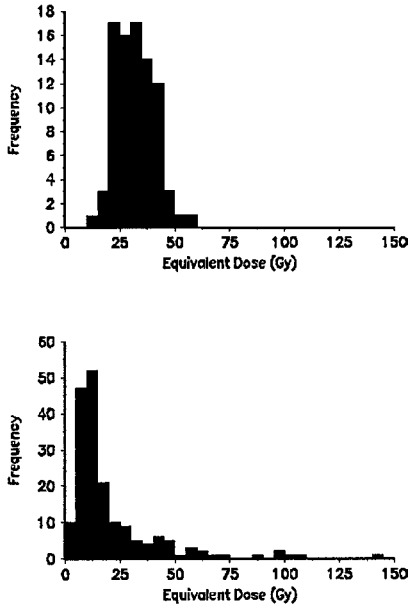


FIGURE 3 Histograms of equivalent dose (ED) values measured from single aliquots of two samples. The first sample has a normal distribution while the second sample has a broad range of ED values indicating that not all the grains in the sample were zeroed at deposition. In this case, the most likely age of deposition is that corresponding to the peak in the histogram at low values of ED.

growth curve is fitted to this regenerated dataset, and the ED is calculated by the intersection of the natural luminescence signal with the growth curve.

For TL measurements, at glow-curve temperatures for which the associated trapped electron lifetimes at ambient temperatures are long compared with the age of the sample, the ED should be independent of temperature. The existence of such an ED plateau is indicative of reliability.

For OSL measurements it is not possible to plot such a plateau since OSL measurements do not provide any information about the thermal stability of the traps from which charge is evicted. It is therefore essential to ensure that only charge from thermally stable traps will be measured during OSL procedures. This is normally achieved by annealing the sample at a fixed temperature for a fixed period of time—a procedure known as preheating—prior to OSL measurement.

C. Single Aliquot Methods

The methods of ED determination described above involve the measurement of many subsamples. Several separate subsamples, or aliquots, are used for measurement of each point on the growth curve, so that between 20 and 60 aliquots are used for each ED determination. An alternative to such multiple aliquot methods is to make all the measurements on a single aliquot. The methods used are very similar to those for multiple aliquots, with both additive dose and regenerative measurements possible.

The advantages of single aliquot methods are that there is no implicit assumption that all aliquots are equivalent, that it becomes practical to make replicate ED determinations on a sample, and that one can alter the size of the sample being analyzed. In the ultimate, one can reduce the sample size to a single mineral grain.

Replicate measurements of the ED using single aliquot methods provide insight into the heterogeneity of the ED from a sample (Fig. 3). For a "simple" sample, a normal distribution is expected, with all aliquots yielding similar ED values. For some materials (for instance, fluvial sands), some grains are exposed to daylight for a limited period of time at deposition, yielding a distribution of grains with different residual signals, only some of which will yield an accurate age. Single aliquot methods can be used to measure the distribution of ED within a sample and, coupled with models of the depositional process, can be used to provide limiting ages.

An additional cause of a broad distribution in ED values can be variation in the annual radiation dose from one grain to another. Fortunately, for most samples this appears to be a relatively small source of variation.

D. Annual Radiation Dose

The annual radiation dose (D in Eq. (1)) to a sample originates from alpha, beta, and gamma radiation. By selecting certain minerals and specific grain size ranges for analysis we can simplify the calculation of the dose rate.

Because of their short range, alpha particles do not penetrate more than about 20 μm into a grain of quartz or feldspar. If we select grains in the range from 4 to 11 μm, they will have received the full alpha dose as well as the beta and gamma contributions.

Alternatively, for large (>100 μm) quartz or feldspar grains it is possible to chemically etch away the outer skin of the grains using hydrofluoric acid and thus remove the alpha contribution. For quartz grains the assumption is normally made that they are free from radioactivity and receive all their beta and gamma dose from the matrix in which they are found. For feldspars the situation is more complex since certain feldspars contain up to 14% potassium by weight. For these grains it is necessary to assess the beta dose originating within the grains. As the grain size increases this internal beta dose becomes increasingly important.

The unit of measurement for the annual radiation dose is the gray (joules deposited per kilogram). In terms of this, alpha particles—because of the high ionization density produced—are not as effective as beta or gamma radiation; the effective full alpha contribution $D'_\alpha = aD_\alpha$, where D_α is the actual alpha dose in grays, and the a-value (a) measured for each sample is typically in the range 0.1 to 0.3. The annual dose for insertion into Eq. (1) is then:

$$D = D'_\alpha + D_\beta + D_\gamma + D_c \qquad (3)$$

where the suffix indicates the type of radiation, with "c" denoting the dose due to cosmic rays. For sediments containing typical concentrations of 1% potassium, 6 ppm thorium, and 2 ppm uranium, D is of the order of 2 grays per 1000 yr for a 200-μm quartz grain.

Allowance needs to be made for the wetness of the sample during burial because the presence of water attenuates the radiation flux received. Uncertainty about average water content is a serious barrier to reducing the error limits on the age to below $\pm$5%.

A variety of methods are used for assessment of radioactivity: neutron activation, chemical analysis, thick-source alpha and beta counting, gamma spectrometry, and beta and gamma dosimetry using special high-sensitivity phosphors (such as calcium sulfate doped with dysprosium, or aluminium oxide doped with carbon). For gamma dosimetry, a capsule of phosphor is buried onsite for about a year. An alternative method of making *in situ* measurements is with a portable gamma spectrometer based around a NaI crystal. Such devices can be used to assess both the total gamma dose rate, and the concentrations of the major radionuclides.

A potential complication in the calculation of the annual dose arises from the possibility that the uranium or thorium decay chains may not be in equilibrium; the most commonly observed cause of disequilibrium is escape of the gas radon-222 which occurs midway in the uranium-238 chain; there may also be preferential leaching of some radioelements in the chain, and the extent to which this occurs may have varied during burial. Thus, disequilibrium can be a problem for two reasons: first, different methods for assessing the total alpha, beta, and gamma dose rates may only look at a specific part of the decay chain and then assume that the remainder of the chain is in equilibrium; and, second, because the extent of disequilibrium may have varied through the period for which the sediment has been buried, causing the annual radiation dose to vary through time. Fortunately, severe disequilibrium is rare.

III. SAMPLE TYPES AND AGE RANGES

A. Baked Clay and Burned Stones

Luminescence dating can reach back to the earliest pottery, at about 10,000 yr ago, and beyond. How much beyond depends on the minerals concerned and the annual radiation dose rate at a specific site. For quartz, the onset of saturation is liable to occur around 50,000 to 100,000 yr ago, and sooner for clay of high radioactivity. TL measurements are particularly applicable in this case since the event being dated is the last heating of the sample in a kiln, oven, or fireplace.

Each sample should be at least 10 mm thick and 30 mm across; six samples per context are desirable. The error limits are usually $\pm$5 to $\pm$10% of the age. An important requirement is that the sample has been buried to a depth of at least 0.3 m and that samples of burial soil are available; if possible, *in situ* radioactivity measurements are also made.

The same age range limitations apply for burned stones as for baked clay; there may be additional difficulties because of heterogeneity of radioactivity and TL sensitivity. Burned flint (and chert) is not plagued in this way and is excellent material for dating; ages of several hundred thousand years have been obtained with it. The main limitation is the sparsity, on palaeolithic sites, of large enough flints that have been sufficiently heated.

B. Authenticity Testing

Luminescence dating has had a very powerful impact on testing authenticity of art ceramics. Error limits of $\pm$25% are typical in this application, but acceptable. The wide error limits arise because of uncertainty as to D_γ and because the allowable sample size is smaller—about 100 mg of powder are obtained by drilling a small hole, 4 mm across by 4 mm deep, in an unobtrusive location. For porcelain, a 3-mm core is extracted and cut into 0.5-mm slices for

measurement. It is also possible to test the authenticity of bronze heads, etc., which have a clay core baked in the casting process.

C. Aeolian Sediments: Loess and Dunes

Aeolian sediments are ideally suited to luminescence dating since one can safely assume that they were exposed to sufficient daylight at deposition to reduce the residual luminescence signal to a low level. Loess, a fine-grained, wind-blown sediment, has been dated extensively in China, the U.S., and Europe. Coastal and desert dunes are also well suited to luminescence dating for the same reason as loess. For loess the grain-size distribution is restricted, and polymineral grains from 4 to 11 μm are routinely used. By contrast, coarse grains between 100 and 300 μm can be obtained from dunes, and grains of quartz or potassium-rich feldspar can be isolated.

The oldest ages are limited by saturation of the luminscence signals. In exceptional circumstances, ages up to 800 ka have been produced, but this is only possible in environments of unusually low radioactivity. Conventionally, the limits are 100,000 yr for quartz and approximately 200,000 yr for feldspars. The youngest ages can be obtained using OSL measurements and are limited by the intrinsic brightness of the minerals concerned and the degree of exposure to daylight at deposition. For well-bleached dunes, ages of a few decades are possible.

Samples should be collected without any exposure to daylight. Sample mass is typically 250 to 500 g but is dependent on the mineralogy and grain-size distribution. Separate samples should also be taken for assessment of water content and radioactivity.

D. Fluvial, Glacial, and Colluvial Sediments

A large variety of sediments are exposed to a limited amount of daylight at deposition, so the exposure varies significantly from one grain to another. For instance, fluvially deposited sands may be well exposed to daylight in shallow rivers in Australia with intense sunlight, but poorly exposed in a deep, turbid, northern European river. Such sediments pose a challenge for luminescence dating methods since they will contain different residual signals at deposition. Included in this list are glacially derived and colluvial sediments.

Single aliquot measurements provide an independent method of assessing whether the sample was sufficiently exposed to daylight at deposition to zero the grains.

SEE ALSO THE FOLLOWING ARTICLES

DOSIMETRY • GEOLOGIC TIME • RADIATION PHYSICS • RADIOCARBON DATING • RADIOMETRIC DATING • STABLE ISOTOPES AS TRACERS OF GLOBAL CYCLES

BIBLIOGRAPHY

Aitken, M. J. (1990). "Science-Based Dating in Archaeology," Longman, London.

Aitken, M. J. (1998). "An Introduction to Optical Dating," Oxford University Press, London.

McKeever, S. W. S. (1985). "Thermoluminescence of Solids," Cam-

Wagner, G. A. (1998). "Age Determination of Young Rocks and Artifacts," Springer-Verlag, Berlin.

Time and Frequency

Michael A. Lombardi
National Institute of Standards and Technology

GLOSSARY

Accuracy Degree of conformity of a measured or calculated value to its definition.
Allan deviation $\sigma_\gamma, (\tau)$ Statistic used to estimate frequency stability.
Coordinated Universal Time (UTC) International atomic time scale used by all major countries.
Nominal frequency Ideal frequency, with zero uncertainty relative to its definition.
Q Quality factor of an oscillator, estimated by dividing the resonance frequency by the resonance width.
Resonance frequency Natural frequency of an oscillator, based on the repetition of a periodic event.
Second Duration of 9,192,631,770 periods of the radiation corresponding to the transition between two hyperfine levels of the ground state of the cesium-133 atom.
Stability Statistical estimate of the frequency or time fluctuations of a signal over a given time interval.
Synchronization Process of setting two or more clocks to the same time.
Syntonization Process of setting two or more oscillators to the same frequency.
Time scale Agreed-upon system for keeping time.

THIS article is an overview of time and frequency technology. It introduces basic time and frequency concepts and describes the devices that produce time and frequency signals and information. It explains how these devices work and how they are measured. Section I introduces the basic concepts of time and frequency and provides some historical background. Section II discusses time and frequency measurements and the specifications used to state the measurement results. Section III discusses time and frequency standards. These devices are grouped into two categories: quartz and atomic oscillators. Section IV discusses time and frequency transfer, or the process of using a clock or frequency standard to measure or set a device at another location.

I. CONCEPTS AND HISTORY

Few topics in science and technology are as relevant as time and frequency. Time and frequency standards and measurements are involved in nearly every aspect of daily life and are a fundamental part of many technologies.

Time and frequency standards supply three basic types of information. The first type, *date* and *time-of-day*,

records when an event happened. Date and time-of-day can also be used to ensure that events are *synchronized*, or happen at the same time. Everyday life is filled with examples of the use of date and time-of-day information. Date information supplied by calendars records when birthdays, anniversaries, and other holidays are scheduled to occur. The time-of-day information supplied by watches and clocks helps keep our lives on schedule. Meeting a friend for dinner at 6 P.M. is a simple example of synchronization. If our watches agree, we should both arrive at about the same time.

Date and time-of-day information has other, more sophisticated uses. Airplanes flying in a formation require synchronized clocks. If one airplane banks or turns at the wrong time, it could result in a collision and loss of life. When a television station broadcasts a network program, it must start broadcasting the network feed at the instant it arrives. If the station and network clocks are not synchronized, part of the program is skipped. Stock market transactions require synchronized clocks so that the buyer and seller can agree on the same price at the same time. A time error of a few seconds could cost the buyer or seller many thousands of dollars. Electric power companies also use synchronization. They synchronize the clocks in their power grids, so they can instantly transfer power to the parts of the grid where it is needed most, and to avoid electrical overload.

The second type of information, *time interval*, is the duration or elapsed time between two events. Our age is simply the time interval since our birth. Most workers are paid for the time interval during which they work, usually measured in hours, weeks, or months. We pay for time interval as well—30 min on a parking meter, a 20-min cab ride, a 5-min long-distance phone call, or a 30-sec radio advertising spot.

The standard unit of time interval is the *second*. However, many applications in science and technology require the measurement of time intervals much shorter than 1 sec, such as *milliseconds* (10^{-3} sec), *microseconds* (10^{-6} sec), *nanoseconds* (10^{-9} sec), and *picoseconds* (10^{-12} sec).

The third type of information, *frequency*, is the rate of a repetitive event. If T is the period of a repetitive event, then the frequency f is its reciprocal, $1/T$. The International System of Units (SI) states that the period should be expressed as seconds (sec), and the frequency should be expressed as hertz (Hz). The frequency of electrical signals is measured in units of kilohertz (kHz), megahertz (MHz), or gigahertz (GHz), where 1 kHz equals 1000 (10^3) events per second, 1 MHz equals 1 million (10^6) events per second, and 1 GHz equals 1 billion (10^9) events per second. Many frequencies are encountered in everyday life. For example, a quartz wristwatch works by counting the oscillations of a crystal whose frequency is 32,768 Hz. When the crystal has oscillated 32,768 times, the watch records that 1 sec has elapsed. A television tuned to channel 7 receives a video signal at a frequency of 175.25 MHz. The station transmits this frequency as closely as possible, to avoid interference with signals from other stations. A computer that processes instructions at a frequency of 1 GHz might connect to the Internet using a T1 line that sends data at a frequency of 1.544 MHz.

Accurate frequency is critical to communications networks. The highest-capacity networks run at the highest frequencies. Networks use groups of oscillators that produce nearly the same frequency, so they can send data at the fastest possible rates. The process of setting multiple oscillators to the same frequency is called *syntonization*.

Of course, the three types of time and frequency information are closely related. As mentioned, the standard unit of time interval is the second. By counting seconds, we can determine the date and the time-of-day. And by counting the events per second, we can measure the frequency.

A. The Evolution of Time and Frequency Standards

All time and frequency standards are based on a *periodic event* that repeats at a constant rate. The device that produces this event is called a *resonator*. In the simple case of a pendulum clock, the pendulum is the resonator. Of course, a resonator needs an energy source before it can move back and forth. Taken together, the energy source and resonator form an *oscillator*. The oscillator runs at a rate called the *resonance frequency*. For example, a clock's pendulum can be set to swing back and forth at a rate of once per second. Counting one complete swing of the pendulum produces a time interval of 1 sec. Counting the total number of swings creates a *time scale* that establishes longer time intervals, such as minutes, hours, and days. The device that does the counting and displays or records the results is called a *clock*. The frequency uncertainty of a clock's resonator relates directly to the timing uncertainty of the clock as shown in Table I.

Throughout history, clock designers have searched for stable resonators. As early as 3500 B.C., time was kept by observing the movement of an object's shadow between sunrise and sunset. This simple clock is called a *sundial*, and the resonance frequency is based on the apparent motion of the sun. Later, water clocks, hourglasses, and calibrated candles allowed dividing the day into smaller units of time. Mechanical clocks first appeared in the early 14th century. Early models used a verge and foliet mechanism for a resonator and had an uncertainty of about 15 min/day ($\cong 1 \times 10^{-2}$).

A timekeeping breakthrough occurred with the invention of the *pendulum clock*, a technology that dominated

TABLE I Relationship of Frequency Uncertainty to Time Uncertainty

Frequency uncertainty	Measurement period	Time uncertainty
$\pm 1.00 \times 10^{-3}$	1 sec	± 1 msec
$\pm 1.00 \times 10^{-6}$	1 sec	± 1 μsec
$\pm 1.00 \times 10^{-9}$	1 sec	± 1 nsec
$\pm 2.78 \times 10^{-7}$	1 hr	± 1 msec
$\pm 2.78 \times 10^{-10}$	1 hr	± 1 μsec
$\pm 2.78 \times 10^{-13}$	1 hr	± 1 nsec
$\pm 1.16 \times 10^{-8}$	1 day	± 1 msec
$\pm 1.16 \times 10^{-11}$	1 day	± 1 μsec
$\pm 1.16 \times 10^{-14}$	1 day	± 1 nsec

timekeeping for several hundred years. Prior to the invention of the pendulum, clocks could not count minutes reliably, but pendulum clocks could count seconds. In the early 1580s, Galileo Galilei observed that a given pendulum took the same amount of time to swing completely through a wide arc as it did a small arc. Galileo wanted to apply this natural periodicity to time measurement and began work on a mechanism to keep the pendulum in motion in 1641, the year before he died. In 1656, the Dutch scientist Christiaan Huygens invented an escapement that kept the pendulum swinging. The uncertainty of Huygens's clock was less than 1 min/day ($\cong 7 \times 10^{-4}$) and later was reduced to about 10 sec/day ($\cong 1 \times 10^{-4}$). The first pendulum clocks were weight-driven, but later versions were powered by springs. In fact, Huygens is often credited with inventing the spring-and-balance wheel assembly still found in some of today's mechanical wristwatches.

Huge advances in accuracy were made by John Harrison, who built and designed a series of clocks in the 1720s that kept time to within fractions of a second per day (parts in 10^6). This performance was not improved upon until the 20th century. Harrison dedicated most of his life to solving the British navy's problem of determining longitude, by attempting to duplicate the accuracy of his land clocks at sea. He built a series of clocks (now known as H1 through H5) in the period from 1730 to about 1770. He achieved his goal with the construction of H4, a clock much smaller than its predecessors, about the size of a large pocket watch. H4 used a spring and balance wheel escapement and kept time within fractions of a second per day during several sea voyages in the 1760s.

The practical performance limit of pendulum clocks was reached in 1921, when W. H. Shortt demonstrated a clock with two pendulums, one a slave and the other a master. The slave pendulum moved the clock's hands and freed the master pendulum of tasks that would disturb its regularity. The pendulums used a battery as their power supply. The Shortt clock kept time to within a few seconds per year ($\cong 1 \times 10^{-7}$) and was used as a primary standard in the United States.

Joseph W. Horton and Warren A. Marrison of Bell Laboratories built the first clock based on a quartz crystal oscillator in 1927. By the 1940s, quartz clocks had replaced pendulums as primary laboratory standards. Quartz crystals resonate at a nearly constant frequency when an electric current is applied. Uncertainties of <100 μsec/day ($\cong 1 \times 10^{-9}$) are possible, and low-cost quartz oscillators are found in electronic circuits and inside nearly every wristwatch and wall clock.

Quartz oscillators still have shortcomings since their resonance frequency depends on the size and shape of the crystal. No two crystals can be precisely alike or produce exactly the same frequency. Quartz oscillators are also sensitive to temperature, humidity, pressure, and vibration. These limitations made them unsuitable for some high-level applications and led to the development of atomic oscillators.

In the 1930s, I. I. Rabi and his colleagues at Columbia University introduced the idea of using an atomic resonance as a frequency. The first atomic oscillator, based on the ammonia molecule, was developed at the National Bureau of Standards (now the National Institute of Standards and Technology) in 1949. A Nobel Prize was awarded in 1989 to Norman Ramsey, Hans Dehmelt, and Wolfgang Paul for their work in atomic oscillator development, and many other scientists have made significant contributions to the technology. Atomic oscillators use the quantized energy levels in atoms and molecules as the source of their resonance frequency. The laws of quantum mechanics dictate that the energies of a bound system, such as an atom, have certain discrete values. An electromagnetic field at a particular frequency can boost an atom from one energy level to a higher one. Or, an atom at a high energy level can drop to a lower level by emitting energy. The resonance frequency (f) of an atomic oscillator is the difference between the two energy levels divided by Planck's constant (h):

$$f = \frac{E_2 - E_1}{h}.$$

The principle underlying the atomic oscillator is that since all atoms of a specific element are identical, they should produce the exact same frequency when they absorb or release energy. In theory, the atom is a perfect pendulum whose oscillations are counted to measure the time interval. Quartz and the three main types of atomic oscillators (rubidium, hydrogen, and cesium) are described in detail in Section III.

Table II summarizes the evolution of time and frequency standards. The uncertainties listed for modern standards

TABLE II The Evolution of Time and Frequency Standards

Standard	Resonator	Date of origin	Timing uncertainty (24 hr)	Frequency uncertainty (24 hr)
Sundial	Apparent motion of sun	3500 B.C.	NA	NA
Verge escapement	Verge and foliet mechanism	14th century	15 min	1×10^{-2}
Pendulum	Pendulum	1656	10 sec	7×10^{-4}
Harrison chronometer (H4)	Pendulum	1759	300 msec	3×10^{-6}
Shortt pendulum	Two pendulums, slave and master	1921	10 msec	1×10^{-7}
Quartz crystal	Quartz crystal	1927	10 μsec	1×10^{-10}
Rubidium gas cell	^{87}Rb resonance (6,834,682,608 Hz)	1958	100 nsec	1×10^{-12}
Cesium beam	^{133}Cs resonance (9,192,631,770 Hz)	1952	1 nsec	1×10^{-14}
Hydrogen maser	Hydrogen resonance (1,420,405,752 Hz)	1960	1 nsec	1×10^{-14}
Cesium fountain	^{133}Cs resonance (9,192,631,770 Hz)	1991	100 psec	1×10^{-15}

represent current (year 2000) devices, and not the original prototypes. Note that the performance of time and frequency standards has improved by 13 orders of magnitude in the past 700 years and by about 9 orders of magnitude in the past 100 years.

B. Time Scales and the International Definition of the Second

The second is one of seven base units in the International System of Units (SI). The base units are used to derive other units of physical quantities. Use of the SI means that physical quantities such as the second and hertz are defined and measured in the same way throughout the world.

There have been several definitions of the SI second. Until 1956, the definition was based on the *mean solar day*, or one revolution of the earth on its axis. The *mean solar second* was defined as 1/86,400 of the mean solar day and provided the basis for several astronomical time scales known as Universal Time (UT).

UT0: The original mean solar time scale, based on the rotation of the earth on its axis. UT0 was first kept with pendulum clocks. When quartz clocks became available, astronomers noticed errors in UT0 due to polar motion and developed the UT1 time scale.

UT1: The most widely used astronomical time scale, UT1 improves upon UT0 by correcting for longitudinal shifts of the observing station due to polar motion. Since the earth's rotational rate is not uniform the uncertainty of UT1 is about 2 to 3 msec per day.

UT2: Mostly of historical interest, UT2 is a smoothed version of UT1 that corrects for deviations in the period of the earth's rotation caused by angular momenta of the earth's core, mantle, oceans and atmosphere.

The *ephemeris second* served as the SI second from 1956 to 1967. The ephemeris second was a fraction of the tropical year, or the interval between the annual vernal equinoxes, which occur on or about March 21. The tropical year was defined as 31,556,925.9747 ephemeris sec. Determining the precise instant of the equinox is difficult, and this limited the uncertainty of Ephemeris Time (ET) to ± 50 msec over a 9-year interval. ET was used mainly by astronomers and was replaced by *Terrestial Time* (TT) in 1984, equal to International Atomic Time (TAI) + 32.184 sec. The uncertainty of TT is ± 10 μsec.

The era of atomic time keeping formally began in 1967, when the SI second was redefined based on the resonance frequency of the cesium atom:

The duration of 9,192,631,770 periods of the radiation corresponding to the transition between two hyperfine levels of the ground state of the cesium-133 atom.

Due to the atomic second, time interval and frequency can now be measured with less uncertainty and more resolution than any other physical quantity. Today, the best time and frequency standards can realize the SI second with uncertainties of $\cong 1 \times 10^{-15}$. Physical realizations of the other base SI units have much larger uncertainties (Table III).

International Atomic Time (TAI) is an atomic time scale that attempts to realize the SI second as closely as possible. TAI is maintained by the Bureau International des Poids et Measures (BIPM) in Sevres, France. The BIPM averages data collected from more than 200 atomic time and frequency standards located at more than 40 laboratories, including the National Institute of Standards and Technology (NIST).

Coordinated Universal Time (UTC) runs at the same rate as TAI. However, it differs from TAI by an integral number of seconds. This difference increases when *leap*

TABLE III Uncertainties of Physical Realizations of the Base SI Units

SI base unit	Physical quantity	Uncertainty
Candela	Luminous intensity	10^{-4}
Mole	Amount of substance	10^{-7}
Kelvin	Thermodynamic temperature	10^{-7}
Ampere	Electric current	10^{-8}
Kilogram	Mass	10^{-8}
Meter	Length	10^{-12}
Second	Time interval	10^{-15}

seconds occur. When necessary, leap seconds are added to UTC on either June 30 or December 31. The purpose of adding leap seconds is to keep atomic time (UTC) within ± 0.9 sec of astronomical time (UT1). Some time codes contain a UT1 correction that can be applied to UTC to obtain UT1.

Leap seconds have been added to UTC at a rate of slightly less than once per year, beginning in 1972. UT1 is currently losing about 700 to 800 msec per year with respect to UTC. This means that atomic seconds are shorter than astronomical seconds and that UTC runs faster than UT1. There are two reasons for this. The first involves the definition of the atomic second, which made it slightly shorter than the astronomical second to begin with. The second reason is that the earth's rotational rate is gradually slowing down and the astronomical second is gradually getting longer. When a positive leap second is added to UTC, the sequence of events is as follows.

23 hr 59 min 59 sec
23 hr 59 min 60 sec
0 hr 0 min 0 sec

The insertion of the leap second creates a minute that is 61 sec long. This "stops" UTC for 1 sec, so that UTI can catch up.

II. TIME AND FREQUENCY MEASUREMENT

Time and frequency measurements follow the conventions used in other areas of metrology. The frequency standard or clock being measured is called the *device under test* (DUT). The measurement compares the DUT to a *standard* or *reference*. The standard should outperform the DUT by a specified ratio, ideally by 10:1. The higher the ratio, the less averaging is required to get valid measurement results.

The test signal for time measurements is usually a pulse that occurs once per second (1 pps). The pulse width and polarity vary from device to device, but TTL levels are commonly used. The test signal for frequency measurements is usually a frequency of 1 MHz or higher, with 5 or 10 MHz being common. Frequency signals are usually sine waves but can be pulses or square waves.

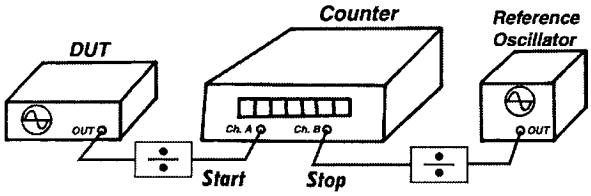


FIGURE 1 Measurement using a time interval counter.

This section examines the two main specifications of time and frequency measurements—*accuracy* and *stability*. It also discusses some instruments used to measure time and frequency.

A. Accuracy

Accuracy is the degree of conformity of a measured or calculated value to its definition. Accuracy is related to the offset from an ideal value. For example, *time offset* is the difference between a measured on-time pulse and an ideal on-time pulse that coincides exactly with UTC. *Frequency offset* is the difference between a measured frequency and an ideal frequency with zero uncertainty. This ideal frequency is called the *nominal frequency*.

Time offset is usually measured with a *time interval counter* (TIC) as shown in Fig. 1. A TIC has inputs for two signals. One signal starts the counter and the other signal stops it. The time interval between the start and the stop signals is measured by counting cycles from the time base oscillator. The resolution of low-cost TICs is limited to the period of their time base. For example, a TIC with a 10-MHz time base oscillator would have a resolution of 100 nsec. More elaborate TICs use interpolation schemes to detect parts of a time base cycle and have a much higher resolution—1-nsec resolution is commonplace, and even 10-psec resolution is available.

Frequency offset can be measured in either the *frequency domain* or the *time domain*. A simple frequency domain measurement involves directly counting and displaying the frequency output of the DUT with a *frequency counter*. The reference for this measurement is either the counter's internal time base oscillator, or an external time base (Fig. 2). The counter's resolution, or the number of digits it can display, limits its ability to measure frequency offset. The frequency offset is determined as

$$f(\text{offset}) = \frac{f_{\text{measured}} - f_{\text{nominal}}}{f_{\text{nominal}}},$$

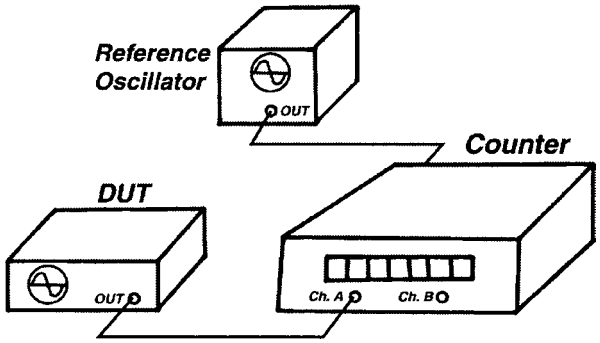


FIGURE 2 Measurement using a frequency counter.

where f_{measured} is the reading from the frequency counter, and f_{nominal} is the frequency labeled on the oscillator's nameplate.

Frequency offset measurements in the time domain involve a *phase comparison* between the DUT and the reference. A simple phase comparison can be made with an oscilloscope (Fig. 3). The oscilloscope will display two sine waves (Fig. 4). The top sine wave represents a signal from the DUT, and the bottom sine wave represents a signal from the reference. If the two frequencies were exactly the same, their phase relationship would not change and both would appear to be stationary on the oscilloscope display. Since the two frequencies are not exactly the same, the reference appears to be stationary and the DUT signal moves. By determining the rate of motion of the DUT signal, we can determine its frequency offset. Vertical lines have been drawn through the points where each sine wave passes through zero. The bottom of the figure shows bars whose width represents the phase difference between the signals. This difference increases or decreases to indicate whether the DUT frequency is high or low with respect to the reference.

Measuring high-accuracy signals with an oscilloscope is impractical, since the phase relationship between signals changes very slowly. More precise phase comparisons can be made with a time interval counter, using a setup similar to Fig. 1. Since frequencies like 5 or 10 MHz are usually involved, *frequency dividers* (shown in Fig. 1) or *frequency mixers* are used to convert the test frequency

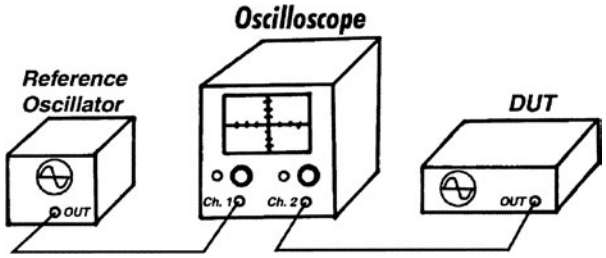


FIGURE 3 Phase comparison using an oscilloscope.

to a lower frequency. Measurements are made from the TIC, but instead of using these measurements directly, we determine the rate of change from reading to reading. This rate of change is called the phase deviation. We can estimate frequency offset as follows, where Δt is the amount of phase deviation, and T is the measurement period:

$$f(\text{offset}) = \frac{-\Delta t}{T}.$$

To illustrate, consider a measurement of $+1$ μsec of phase deviation over a measurement period of 24 hr. The unit used for measurement period (hr) must be converted to the unit used for phase deviation (μsec). The equation becomes

$$f(\text{offset}) = \frac{-\Delta t}{T} = \frac{1\ \mu\text{sec}}{86{,}400{,}000{,}000\ \mu\text{sec}}$$
$$= -1.16 \times 10^{-11}.$$

As shown, a device that accumulates 1 μsec of phase deviation/day has a frequency offset of about -1.16×10^{-11} with respect to the reference.

Dimensionless frequency offset values can be converted to units of frequency (Hz) if the nominal frequency is known. To illustrate this, consider an oscillator with a nominal frequency of 5 MHz and a frequency offset of $+1.16 \times 10^{-11}$. To find the frequency offset in hertz, multiply the nominal frequency by the offset:

$$(5 \times 10^6)(+1.16 \times 10^{-11}) = 5.80 \times 10^{-5}$$
$$= +0.0000580 \text{ Hz}.$$

Then add the offset to the nominal frequency to get the actual frequency:

$$5{,}000{,}000 \text{ Hz} + 0.0000580 \text{ Hz}$$
$$= 5{,}000{,}000.0000580 \text{ Hz}.$$

B. Stability

Stability indicates how well an oscillator can produce the same time or frequency offset over a given period of time. It does not indicate whether the time or frequency is "right" or "wrong" but only whether it stays the same. In contrast, accuracy indicates how well an oscillator has been set on time or set on frequency. To understand this difference, consider that a stable oscillator that needs adjustment might produce a frequency with a large offset. Or an unstable oscillator that was just adjusted might temporarily produce a frequency near its nominal value. Figure 5 shows the relationship between accuracy and stability.

Stability is defined as the statistical estimate of the frequency or time fluctuations of a signal over a given time

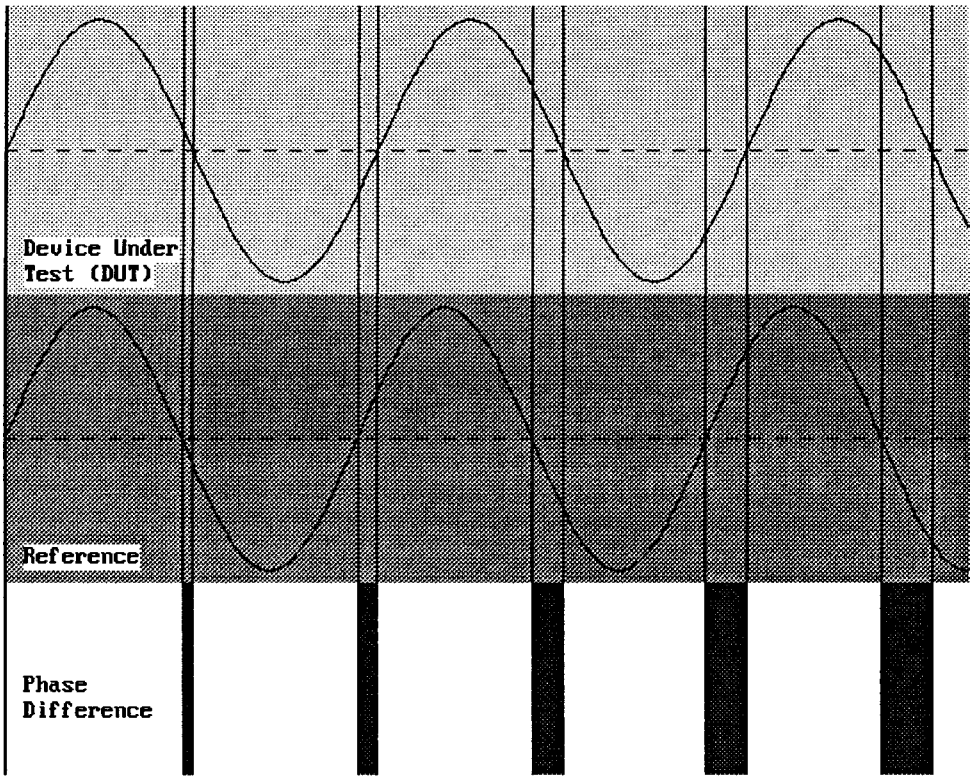


FIGURE 4 Two sine waves with a changing phase relationship.

interval. These fluctuations are measured with respect to a mean frequency or time offset. *Short-term* stability usually refers to fluctuations over intervals less than 100 sec. *Long-term* stability can refer to measurement intervals greater than 100 sec but usually refers to periods longer than 1 day.

Stability estimates can be made in either the frequency domain or the time domain, but time domain estimates are more common and are discussed in this section. To estimate frequency stability in the time domain, we can start with a series of phase measurements. The phase measurements are nonstationary, since they contain a trend contributed by the frequency offset. With nonstationary data, the mean and variance never converge to any particular values. Instead, there is a moving mean that changes each time we add a measurement.

For these reasons, a nonclassical statistic is often used to estimate stability in the time domain. This statistic is sometimes called the *Allan variance*, but since it is the square root of the variance, its proper name is the *Allan deviation*. The equation for the Allan deviation is

$$\sigma_y(\tau) = \sqrt{\frac{1}{2(M-1)} \sum_{i=1}^{M-1} (y_{i+1} - y_i)^2},$$

where y_i is a set of frequency offset measurements that consists of individual measurements, y_1, y_2, y_3, and so on,

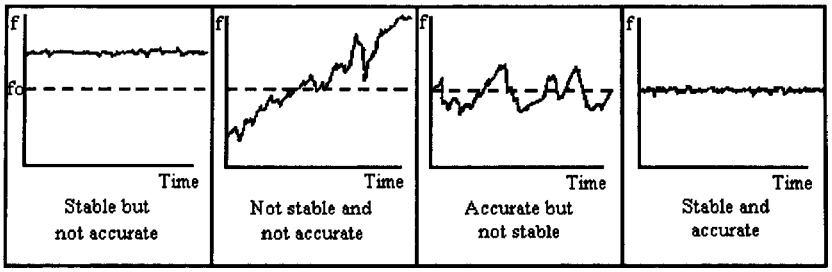


FIGURE 5 The relationship between accuracy and stability.

TABLE IV Using Phase Measurements to Estimate Stability

Phase measurement (nsec), x_i	Phase deviation (nsec), Δt	Frequency offset $\Delta t/\tau(y_i)$	First difference $(y_{i+1} - y_i)$	First difference squared $(y_{i+1} - y_i)^2$
3321.44	(—)	(—)	(—)	(—)
3325.51	4.07	4.07×10^{-9}	(—)	(—)
3329.55	4.04	4.04×10^{-9}	-3×10^{-11}	9×10^{-22}
3333.60	4.05	4.05×10^{-9}	$+1 \times 10^{-11}$	1×10^{-22}
3337.65	4.05	4.06×10^{-9}	$+2 \times 10^{-11}$	4×10^{-22}
3341.69	4.04	4.04×10^{-9}	-2×10^{-11}	4×10^{-22}
3345.74	4.05	4.05×10^{-9}	$+1 \times 10^{-11}$	1×10^{-22}
3349.80	4.06	4.06×10^{-9}	$+1 \times 10^{-11}$	1×10^{-22}
3353.85	4.05	4.05×10^{-9}	-1×10^{-11}	1×10^{-22}
3357.89	4.04	4.04×10^{-9}	-1×10^{-11}	1×10^{-22}

M is the number of values in the y_i series, and the data are equally spaced in segments τ seconds long. Or

$$\sigma_y(\tau) = \sqrt{\frac{1}{2(N-2)\tau^2} \sum_{i=1}^{N-2} [x_{i+2} - 2x_{i+1} + x_i]^2},$$

where x_i is a set of phase measurements in time units that consists of individual measurements, x_1, x_2, x_3, and so on, N is the number of values in the x_i series, and the data are equally spaced in segments τ seconds long.

Table IV shows how the Allan deviation is calculated. The left column contains a series of phase measurements recorded once per second ($\tau = 1$ sec) in units of nanoseconds. These measurements have a trend; note that each value in the series is larger than the previous value. By subtracting pairs of values, we remove the trend and obtain the phase deviations (Δt) shown in the second column. The third column divides the phase deviation (Δt) by τ to get the frequency offset values, or the y_i data series. The last two columns show the first differences of the y_i and the squares of the first differences.

Since the sum of the squares equals 2.2×10^{-21}, the frequency stability using the first equation (at $\tau = 1$ sec) is

$$\sigma_y(\tau) = \sqrt{\frac{2.2 \times 10^{-21}}{2(9-1)}} = 1.17 \times 10^{-11}.$$

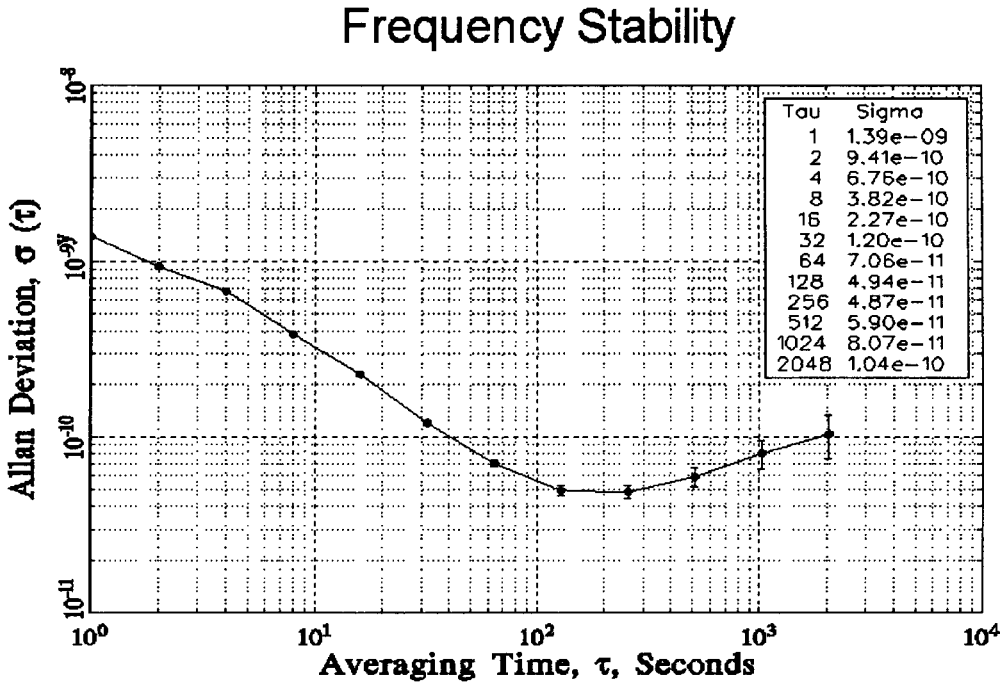


FIGURE 6 A graph of frequency stability.

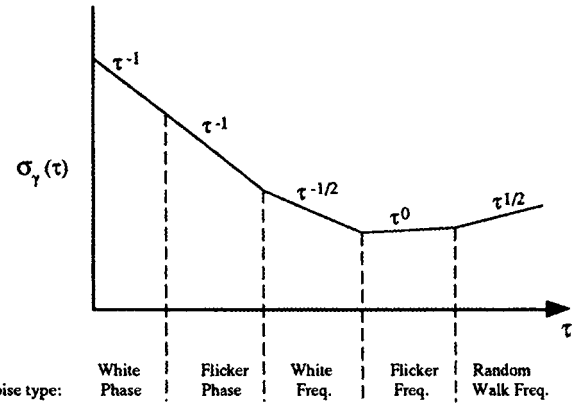


FIGURE 7 Using a frequency stability graph to identify noise types.

A graph of the Allan deviation is shown in Fig. 6. It shows the stability of the device improving as the averaging period (τ) gets longer, since some noise types can be removed by averaging. At some point, however, more averaging no longer improves the results. This point is called the *noise floor*, or the point where the remaining noise consists of nonstationary processes like aging or random walk. The device measured in Fig. 6 is has a noise floor of $\cong 5 \times 10^{-11}$ at $\tau = 100$ sec.

Five noise types are commonly discussed in the time and frequency literature: *white phase*, *flicker phase*, *white frequency*, *flicker frequency*, and *random walk frequency*. The slope of the Allan deviation line can identify the amount of averaging needed to remove these noise types (Fig. 7). Note that the Allan deviation does not distinguish between white phase noise and flicker phase noise. Several other statistics are used to estimate stability and identify noise types for various applications (Table V).

C. Uncertainty Analysis

Time and frequency metrologists must often perform an uncertainty analysis when they calibrate or measure a device. The uncertainty analysis states the measurement error with respect to a national or international standard, such as UTC(NIST) or UTC. Two simple ways to estimate measurement uncertainty are discussed here. Both use the concepts of accuracy and stability discussed above.

One common type of uncertainty analysis involves making multiple measurements and showing that a single measurement will probably fall within a stated range of values. The standard deviation (or an equivalent statistic) is usually both added and subtracted from the mean to form the upper and lower bounds of the range. The stated probability that a given measurement will fall within this range is usually 1σ (68.3%) or 2σ (95.4%).

In time and frequency metrology, the mean value is usually the accuracy (mean time or mean frequency offset), and the deviation in the mean is usually calculated using one of the statistics listed in Table IV. For example, if a device has a frequency offset of 2×10^{-9} and a 2σ stability of 2×10^{-10}, there is a 95.4% probability that the frequency offset will be between 1.8 and 2.2 parts in 10^9.

The second type of uncertainty analysis involves adding the *systematic* and *statistical* uncertainties to find the combined uncertainty. For example, consider a time signal received from a radio station where the mean path delay is measured as 9 msec (time offset), and the deviation in the path delay is measured as 0.5 msec (stability). In this example, 9 msec is the systematic uncertainty and 0.5 msec is the statistical uncertainty. For some applications, it is convenient to simply add the two numbers together and state the combined uncertainty as <10 msec.

III. TIME AND FREQUENCY STANDARDS

The stability of time and frequency standards is closely related to their quality factor, or Q. The Q of an oscillator is its resonance frequency divided by its resonance width. The resonance frequency is the natural frequency of the oscillator. The resonance width is the range of possible values where the oscillator will run. A high-Q resonator will not oscillate at all unless it is near its resonance frequency. Obviously a high resonance frequency and a narrow resonance width are both advantages when seeking a high Q.

TABLE V Statistics Used to Estimate Time and Frequency Stability and Noise Types

Name	Mathematical notation	Description
Allan deviation	$\sigma_y(\tau)$	Estimates frequency stability. Particularly suited for intermediate to long-term measurements.
Modified Allan deviation	MOD $\sigma_y(\tau)$	Estimates frequency stability. Unlike the normal Allan deviation, it can distinguish between white and flicker phase noise, which makes it more suitable for short-term stability estimates.
Time deviation	$\sigma_x(\tau)$	Used to measure time stability. Clearly identifies both white and flicker phase noise, the noise types of most interest when measuring time or phase.
Total deviation	$\sigma_{y,\,\mathrm{TOTAL}}(\tau)$	Estimates frequency stability. Particularly suited for long-term estimates where τ exceeds 10% of the total data sample.

TABLE VI Summary of Oscillator Types

Oscillator type	Quartz		Rubidium	Commercial cesium beam	Hydrogen maser
	TCXO	OCXO			
Q	10^4 to 10^6	3.2×10^6 (5 MHz)	10^7	10^8	10^9
Resonance frequency	Various	Various	6.834682608 GHz	9.192631770 GHz	1.420405752 GHz
Leading cause of failure	None	None	Rubidium lamp (15 years +)	Cesium beam tube (3 to 25 years)	Hydrogen depletion (7 years +)
Stability, $\sigma_y(\tau)$, $\tau = 1$ sec	1×10^{-8} to 1×10^{-9}	1×10^{-12}	5×10^{-11} to 5×10^{-12}	5×10^{-11} to 5×10^{-12}	1×10^{-12}
Noise floor, $\sigma_y(\tau)$	1×10^{-9} ($\tau = 1$ to 10^2 sec)	1×10^{-12} ($\tau = 1$ to 10^2 sec)	1×10^{-12} ($\tau = 10^3$ to 10^5 sec)	1×10^{-14} ($\tau = 10^5$ to 10^7 sec)	1×10^{-15} ($\tau = 10^3$ to 10^5 sec)
Aging/year	5×10^{-7}	5×10^{-9}	1×10^{-10}	None	$\cong 1 \times 10^{-13}$
Frequency offset after warm-up	1×10^{-6}	1×10^{-8} to 1×10^{-10}	5×10^{-10} to 5×10^{-12}	5×10^{-12} to 1×10^{-14}	1×10^{-12} to 1×10^{-13}
Warm-up period	<10 sec to 1×10^{-6}	<5 min to 1×10^{-8}	<5 min to 5×10^{-10}	30 min to 5×10^{-12}	24 hr to 1×10^{-12}

Generally speaking, the higher the Q, the more stable the oscillator, since a high Q means that an oscillator will stay close to its natural resonance frequency.

This section discusses quartz oscillators, which achieve the highest Q of any mechanical-type device. It then discusses oscillators with much higher Q factors, based on the atomic resonance of rubidium, hydrogen, and cesium. The performance of each type of oscillator is summarized in Table VI.

A. Quartz Oscillators

Quartz crystal oscillators are by far the most common time and frequency standard. An estimated 2 billion (2×10^9) quartz oscillators are manufactured annually. Most are small devices built for wristwatches, clocks, and electronic circuits. However, they are also found inside test and measurement equipment, such as counters, signal generators, and oscilloscopes, and interestingly enough, inside every atomic oscillator.

A quartz crystal inside the oscillator is the resonator. It can be made of natural or synthetic quartz, but all modern devices use synthetic quartz. The crystal strains (expands or contracts) when a voltage is applied. When the voltage is reversed, the strain is reversed. This is known as the *piezoelectric effect*. Oscillation is sustained by taking a voltage signal from the resonator, amplifying it, and feeding it back to the resonator. The rate of expansion and contraction is the resonance frequency and is determined by the cut and size of the crystal. The output frequency of a quartz oscillator is either the fundamental resonance or a multiple of the resonance, called an *overtone frequency*. Most high-stability units use either the third or the fifth overtone to achieve a high Q. Overtones higher than fifth are rarely used because they make it harder to tune the device to the desired frequency. A typical Q for a quartz oscillator ranges from 10^4 to 10^6. The maximum Q for a high-stability quartz oscillator can be estimated as $Q = 16$ million$/f$, where f is the resonance frequency in megahertz.

Environmental changes such as temperature, humidity, pressure, and vibration can change the resonance frequency of a quartz crystal, and there are several designs that reduce the environmental problems. The *oven-controlled crystal oscillator* (OCXO) encloses the crystal in a temperature-controlled chamber called an oven. When an OCXO is turned on, it goes through a "warm-up" period while the temperatures of the crystal resonator and its oven stabilize. During this time, the performance of the oscillator continuously changes until it reaches its normal operating temperature. The temperature within the oven, then remains constant, even when the outside temperature varies. An alternate solution to the temperature problem is the *temperature-compensated crystal oscillator* (TCXO). In a TCXO, the signal from a temperature sensor generates a correction voltage that is applied to a voltage-variable reactance, or varactor. The varactor then produces a frequency change equal and opposite to the frequency change produced by temperature. This technique does not work as well as oven control but is less expensive. Therefore, TCXOs are used when high stability over a wide temperature range is not required.

Quartz oscillators have excellent short-term stability. An OCXO might be stable ($\sigma_y \tau$, at $\tau = 1$ sec) to 1×10^{-12}. The limitations in short-term stability are due mainly to noise from electronic components in the oscillator circuits. Long-term stability is limited by *aging*, or a change in frequency with time due to internal changes in the oscillator. Aging is usually a nearly linear change in the resonance frequency that can be either positive or negative, and occasionally, a reversal in aging direction occurs. Aging has many possible causes including a buildup of foreign material on the crystal, changes in the oscillator circuitry, or changes in the quartz material or crystal structure. A

high-quality OCXO might age at a rate of $<5 \times 10^{-9}$ per year, while a TCXO might age 100 times faster.

Due to aging and environmental factors such as temperature and vibration, it is hard to keep even the best quartz oscillators within 1×10^{-10} of their nominal frequency without constant adjustment. For this reason, atomic oscillators are used for applications that require higher long-term accuracy and stability.

B. Rubidium Oscillators

Rubidium oscillators are the lowest priced members of the atomic oscillator family. They operate at 6,834,682,608 Hz, the resonance frequency of the rubidium atom (^{87}Rb), and use the rubidium frequency to control the frequency of a quartz oscillator. A microwave signal derived from the crystal oscillator is applied to the ^{87}Rb vapor within a cell, forcing the atoms into a particular energy state. An optical beam is then pumped into the cell and is absorbed by the atoms as it forces them into a separate energy state. A photo cell detector measures how much of the beam is absorbed and tunes a quartz oscillator to a frequency that maximizes the amount of light absorption. The quartz oscillator is then locked to the resonance frequency of rubidium, and standard frequencies are derived and provided as outputs (Fig. 8).

Rubidium oscillators continue to get smaller and less expensive, and offer perhaps the best price/performance ratio of any oscillator. Their long-term stability is much better than that of a quartz oscillator and they are also smaller, more reliable, and less expensive than cesium oscillators.

The Q of a rubidium oscillator is about 10^7. The shifts in the resonance frequency are caused mainly by collisions of the rubidium atoms with other gas molecules. These shifts limit the long-term stability. Stability ($\sigma_y\tau$, at $\tau = 1$ sec) is typically 1×10^{-11}, and about 1×10^{-12} at 1 day. The frequency offset of a rubidium oscillator ranges from 5×10^{-10} to 5×10^{-12} after a warm-up period of a few minutes, so they meet the accuracy requirements of most applications without adjustment.

C. Cesium Oscillators

Cesium oscillators are *primary frequency standards* since the SI second is defined using the resonance frequency of the cesium atom (^{133}Cs), which is 9,192,631,770 Hz. A properly working cesium oscillator should be close to its nominal frequency without adjustment, and there should be no change in frequency due to aging.

Commercially available oscillators use *cesium beam* technology. Inside a cesium oscillator, ^{133}Cs atoms are heated to a gas in an oven. Atoms from the gas leave the oven in a high-velocity beam that travels through a vacuum tube toward a pair of magnets. The magnets serve as a gate that allows only atoms of a particular magnetic energy state to pass into a microwave cavity, where they are exposed to a microwave frequency derived from a quartz oscillator. If the microwave frequency matches the resonance frequency of cesium, the cesium atoms will change their magnetic energy state.

The atomic beam then passes through another magnetic gate near the end of the tube. Those atoms that changed their energy state while passing through the microwave cavity are allowed to proceed to a detector at the end of the tube. Atoms that did not change state are deflected away from the detector. The detector produces a feedback signal that continually tunes the quartz oscillator in a way that maximizes the number of state changes so that the greatest number of atoms reaches the detector. Standard output frequencies are derived from the locked quartz oscillator (Fig. 9).

The Q of a commercial cesium standard is a few parts in 10^8. The beam tube is typically <0.5 m in length, and the atoms travel at velocities of >100 m per second inside the tube. This limits the observation time to a few

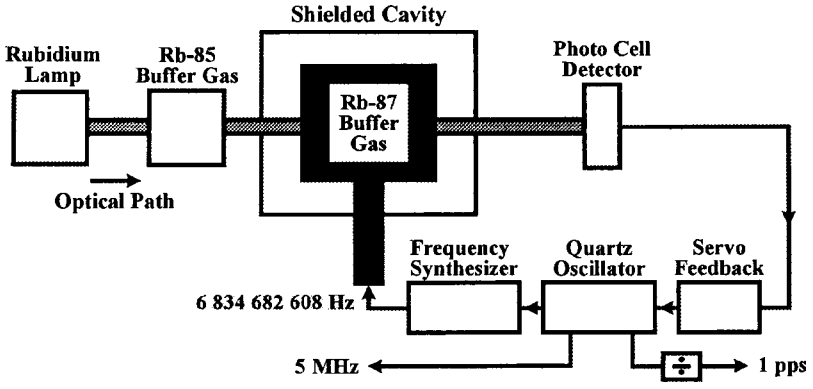


FIGURE 8 Rubidium oscillator.

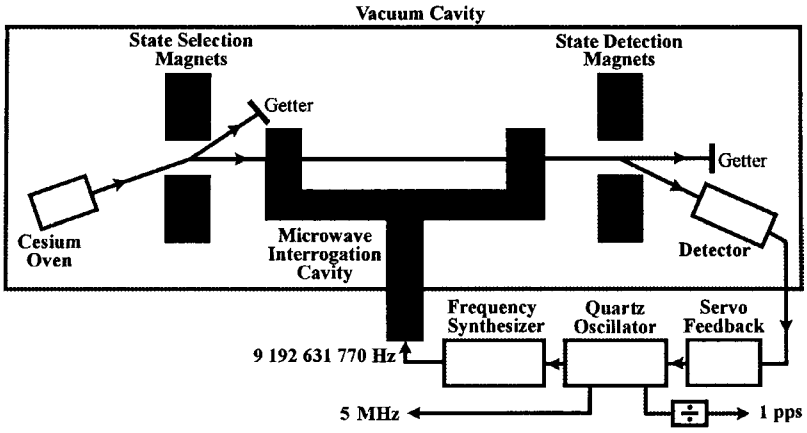


FIGURE 9 Cesium beam oscillator.

milliseconds, and the resonance width to a few hundred hertz. Stability ($\sigma_y\tau$, at $\tau = 1$ sec) is typically 5×10^{-12} and reaches a noise floor near 1×10^{-14} at about 1 day, extending out to weeks or months. The frequency offset is typically near 1×10^{-12} after a warm-up period of 30 min.

The current state-of-the-art in cesium technology is the *cesium fountain* oscillator, named after its fountain-like movement of cesium atoms. A cesium fountain named NIST-F1 serves as the primary standard of time and frequency for the United States.

A cesium fountain works by releasing a gas of cesium atoms into a vacuum chamber. Six infrared laser beams are directed at right angles to each other at the center of the chamber. The lasers gently push the cesium atoms together into a ball. In the process of creating this ball, the lasers slow down the movement of the atoms and cool them to temperatures a few thousandths of a degree above absolute zero. This reduces their thermal velocity to a few centimeters per second.

Two vertical lasers gently toss the ball upward and then all of the lasers are turned off. This little push is just enough to loft the ball about a meter high through a microwave-filled cavity. Under the influence of gravity, the ball then falls back down through the microwave cavity. The round trip up and down through the microwave cavity lasts for about 1 sec and is limited only by the force of gravity pulling the atoms to the ground. During the trip, the atomic states of the atoms might or might not be altered as they interact with the microwave signal. When their trip is finished, another laser is pointed at the atoms. Those atoms whose states were altered by the microwave signal emit photons (a state known as *fluorescence*) that are counted by a detector. This process is repeated many times while the microwave signal in the cavity is tuned to different frequencies. Eventually, a microwave frequency is found that alters the states of most of the cesium atoms and maximizes their fluorescence. This frequency is the cesium resonance (Fig. 10).

The Q of a cesium fountain is about 10^{10}, or about 100 times higher than a traditional cesium beam. Although the resonance frequency is the same, the resonance width is much narrower (<1 Hz), due to the longer observation times made possible by the combination of laser cooling and the fountain design. The combined frequency uncertainty of NIST-F1 is estimated at $<2 \times 10^{-15}$.

D. Hydrogen Masers

The *hydrogen maser* is the most elaborate and expensive commercially available frequency standard. The word *maser* is an acronym that stands for microwave amplification by stimulated emission of radiation. Masers operate at the resonance frequency of the hydrogen atom, which is 1,420,405,752 Hz.

A hydrogen maser works by sending hydrogen gas through a magnetic gate that allows only atoms in certain energy states to pass through. The atoms that make it through the gate enter a storage bulb surrounded by a tuned, resonant cavity. Once inside the bulb, some atoms drop to a lower energy level, releasing photons of microwave frequency. These photons stimulate other atoms to drop their energy level, and they in turn release additional photons. In this manner, a self-sustaining microwave field builds up in the bulb. The tuned cavity around the bulb helps to redirect photons back into the system to keep the oscillation going. The result is a microwave signal that is locked to the resonance frequency of the hydrogen atom and that is continually emitted as long as new atoms are fed into the system. This signal keeps a quartz crystal oscillator in step with the resonance frequency of hydrogen (Fig. 11).

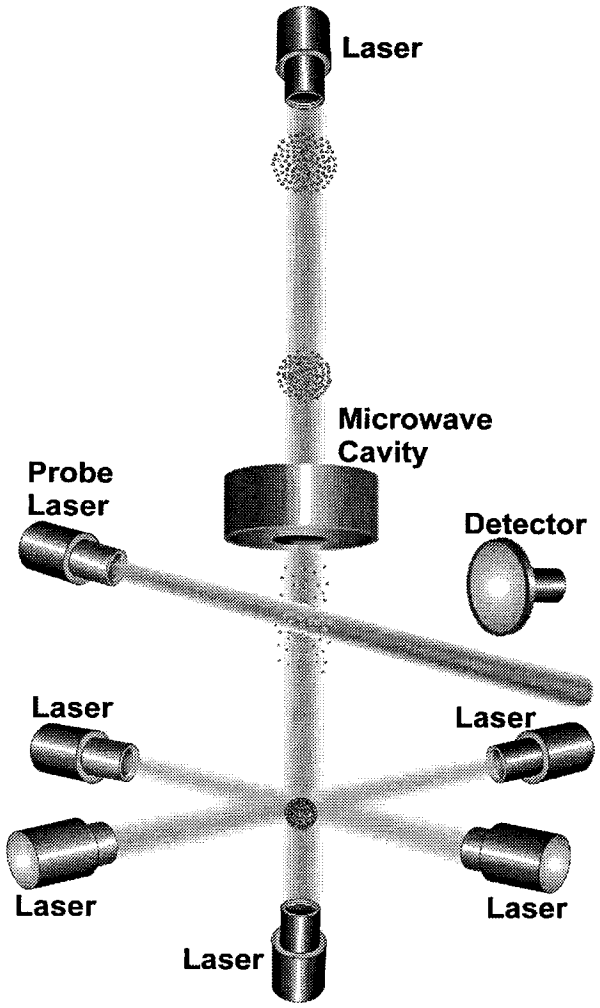


FIGURE 10 Cesium fountain oscillator.

The resonance frequency of hydrogen is much lower than that of cesium, but the resonance width of a hydrogen maser is usually just a few hertz. Therefore, the Q is about 10^9, or at least one order of magnitude better than a commercial cesium standard. As a result, the short-term stability is better than a cesium standard for periods out to a few days—typically $<1 \times 10^{-12}$ ($\sigma_y \tau$, at $\tau = 1$ sec) and reaching a noise floor of $\cong 1 \times 10^{-15}$ after about 1 hr. However, when measured for more than a few days or weeks, a hydrogen maser might fall below a cesium oscillator's performance. The stability decreases because of changes in the cavity's resonance frequency over time.

E. Future Standards

Research conducted at NIST and other laboratories should eventually lead to frequency standards that are far more stable than current devices. Future standards might use the resonance frequency of trapped, electrically charged ions. Trapping ions and suspending them in a vacuum allows them to be isolated from disturbing influences and observed for periods of 100 sec or longer. Much of this work has been based on the mercury ion ($^{199}Hg^{+}$), since its resonance frequency in the microwave realm is about 40.5 GHz, or higher than that of other atoms appropriate for this trapping technique. With a resonance width of 10 mHz or less, the Q of a mercury ion standard can reach 10^{12}.

The most promising application of trapped ions is their use in optical frequency standards. These devices use ion traps that resonate at optical, rather than microwave frequencies. The resonance frequency of these devices is about 10^{15} Hz; for example, the $^{199}Hg^{+}$ ion has an optical wavelength of just 282 nm. Although long observation times are difficult with this approach, experiments have shown that a resonance width of 1 Hz might eventually be possible. This means that the Q of an optical frequency standard could reach 10^{15}, several orders of magnitude higher than the best microwave experiments.

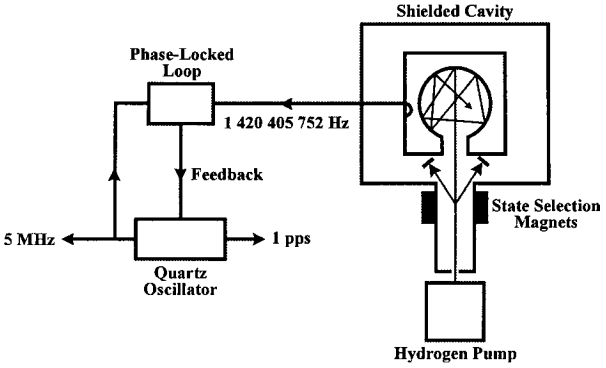


FIGURE 11 Hydrogen maser oscillator.

IV. TIME AND FREQUENCY TRANSFER

Many applications require clocks or oscillators at different locations to be set to the same time (*synchronization*) or the same frequency (*syntonization*). *Time and frequency transfer* techniques are used to compare and adjust clocks and oscillators at different locations. Time and frequency transfer can be as simple as setting your wristwatch to an audio time signal or as complex as controlling the frequency of oscillators in a network to parts in 10^{13}.

Time and frequency transfer can use signals broadcast through many different media, including coaxial cables, optical fiber, radio signals (at numerous places in the spectrum), telephone lines, and the Internet. Synchronization requires both an on-time pulse and a time code.

TABLE VII Summary of Time and Frequency Transfer Signals and Methods

Signal or link	Receiving equipment	Time uncertainty (24 hr)	Frequency uncertainty (24 hr)
Dial-up computer time service	Computer, software, modem, and phone line	<15 msec	NA
Network time service	Computer, software, and Internet connection	<1 sec	NA
HF radio (3 to 30 MHz)	HF receiver	1 to 20 msec	10^{-6} to 10^{-9}
LF radio (30 to 300 kHz)	LF receiver	1 to 100 μsec	10^{-10} to 10^{-12}
GPS one-way	GPS receiver	<50 nsec	$\cong 10^{-13}$
GPS common-view	GPS receiver, tracking schedule (single channel only), data link	<10 nsec	$<1 \times 10^{-13}$
GPS carrier phase	GPS carrier phase tracking receiver, orbital data for postprocessing corrections, data link	<50 nsec	$<1 \times 10^{-14}$
Two-way satellite	Receiving equipment, transmitting equipment, data link	<1 nsec	$<1 \times 10^{-14}$

Syntonization requires extracting a stable frequency from the broadcast, usually from the carrier frequency or time code.

This section discusses both the fundamentals of time and frequency transfer and the radio and network signals used. Table VII provides a summary.

A. Fundamentals of Time and Frequency Transfer

The largest contributor to time transfer uncertainty is *path delay*, or the signal delay between the transmitter and the receiver. For example, consider a radio signal broadcast over a 1000-km path. Since radio signals travel at the speed of light ($\cong$3.3 μsec/km), we can calibrate the path by estimating the path delay as 3.3 msec and applying a 3.3-msec correction to our measurement. The more sophisticated time transfer systems are self-calibrating and automatically correct for path delay.

Path delay is not important to frequency transfer systems, since on-time pulses are not required. Instead, frequency transfer requires only a stable path where the delays remain relatively constant. The three basic types of time and frequency transfer methods are described below.

1. One-Way Method

This is the simplest and most common way to transfer time and frequency information. Information is sent from a transmitter to a receiver and is delayed by the path through the medium (Fig. 12). To get the best results, the user must estimate τ_{ab} and calibrate the path to compensate for the delay. Of course, for many applications the path delay is simply ignored. For example, if our goal is simply to synchronize a computer clock within 1 sec of UTC, there is no need to worry about a 100-msec delay through a network.

More sophisticated one-way transfer systems estimate and remove all or part of the τ_{ab} delay. This is usually

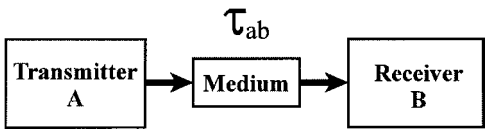


FIGURE 12 One-way transfer.

done in one of two ways. The first way is to estimate τ_{ab} and send the time out early by this amount. For example, if τ_{ab} is at least 20 msec for all users, the time can be sent 20 msec early. This advancement of the timing signal will remove at least some of the delay for all users.

A better technique is to compute τ_{ab} and to apply a correction to the broadcast. A correction for τ_{ab} can be computed if the position of both the transmitter and the receiver are known. If the transmitter is stationary, a constant can be used for the transmitter position. If the transmitter is moving (a satellite, for example), it must broadcast its position in addition to broadcasting time. The Global Positioning System provides the best of both worlds—each satellite broadcasts its position and the receiver can use coordinates from multiple satellites to compute its own position.

One-way time transfer systems often include a *time code* so that a clock can be set to the correct time-of-day. Most time codes contain the UTC hour, minute, and second. Some contain date information, a UT1 correction, and advance warning of daylight savings time and leap seconds.

2. Common-View Method

The common-view method involves a single reference transmitter (R) and two receivers (A and B). The transmitter is in common view of both receivers. Both receivers compare the simultaneously received signal to their local clock and record the data. Receiver A receives the signal over the path τ_{ra} and compares the reference to its local clock (R − Clock A). Receiver B receives the signal over

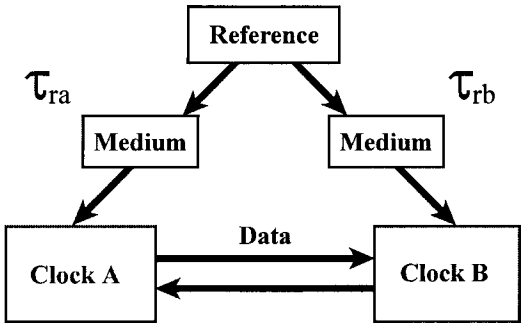


FIGURE 13 Common-view transfer.

the path τ_{rb} and records (R − Clock B). The two receivers then exchange and difference the data (Fig. 13).

Common-view directly compares two time and frequency standards. Errors from the two paths (τ_{ra} and τ_{rb}) that are common to the reference cancel out, and the uncertainty caused by path delay is nearly eliminated. The result of the measurement is (Clock A − Clock B) − ($\tau_{ra} - \tau_{rb}$).

3. Two-Way Method

The two-way method requires two users to both transmit and receive through the same medium at the same time. Sites A and B simultaneously exchange time signals through the same medium and compare the received signals with their own clocks. Site A records A − (B + τ_{ba}) and site B records B − (A + τ_{ab}), where τ_{ba} is the path delay from A to B, and τ_{ab} is the path delay from A to B. The difference between these two sets of readings produces 2(A − B) − ($\tau_{ba} - \tau_{ab}$). Since the path is reciprocal ($\tau_{ab} = \tau_{ba}$), the path delay cancels out of the equation (Fig. 14).

The two-way method is used for international comparisons of time standards using spread spectrum radio signals at C- or Ku-band frequencies, and a geostationary satellite as a transponder. The stability of these comparisons is usually <500 psec ($\sigma_x \tau$, at $\tau = 1$ sec), or $< 1 \times 10^{-14}$ for frequency, even when the clocks are separated by thousands of kilometers.

The two-way method is also used in telecommunications networks where transmission of a signal can be done in software. Some network and telephone time signals use a variation of two-way, called the *loop-back* method. Like

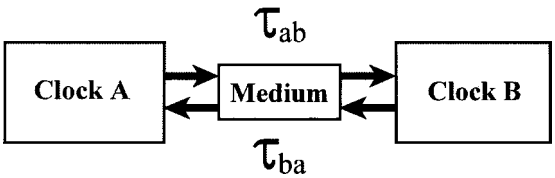


FIGURE 14 Two-way transfer.

the two-way method, the loop-back method requires both users to transmit and receive, but not at the same time. For example, a signal is sent from the transmitter (A) to the receiver (B) over the path τ_{ab}. The receiver (B) then echoes or reflects the signal back to the transmitter (A) over the path τ_{ba}. The transmitter then adds the two path delays ($\tau_{ab} + \tau_{ba}$) to obtain the round-trip delay and divides this number by 2 to estimate the one-way path delay. The transmitter then advances the next time signal by the estimated one-way delay. Since users do not transmit and receive at the same time, the loop-back method has larger uncertainties than the two-way method. A reciprocal path cannot be assumed, since we do not know if the signal from A to B traveled the same path as the signal from B to A.

B. Radio Time and Frequency Transfer Signals

There are many types of radio receivers designed to receive time and frequency information. Radio clocks come in several different forms. Some are tabletop or rack-mount devices with a digital time display and a computer interface. Others are available as cards that plug directly into a computer.

The uncertainty of a radio time transfer system consists of the uncertainty of the received signal, plus delays in the receiving equipment. For example, there is cable delay between the antenna and the receiver. There are equipment delays introduced by hardware, and processing delays introduced by software. These delays must be calibrated to get the best results. When doing frequency transfer, equipment delays can be ignored if they remain relatively constant.

The following sections look at the three types of radio signals most commonly used for time and frequency transfer—high frequency (HF), low frequency (LF), and Global Positioning System (GPS) satellite signals.

1. HF Radio Signals (Including WWV and WWVH)

High-frequency (HF) radio broadcasts occupy the radio spectrum from 3 to 30 MHz. These signals are commonly used for time and frequency transfer at moderate performance levels. Some HF broadcasts provide audio time announcements and digital time codes. Other broadcasts simply provide a carrier frequency for use as a reference.

HF time and frequency stations (Table VIII) include NIST radio stations WWV and WWVH. WWV is located near Fort Collins, Colorado, and WWVH is on the island of Kauai, Hawaii. Both stations broadcast continuous time and frequency signals on 2.5, 5, 10, and 15 MHz, and WWV also broadcasts on 20 MHz. All frequencies carry the same program, and at least one frequency should be usable at all times. The stations can also be heard by

TABLE VIII HF Time and Frequency Broadcast Stations

Call sign	Country	Frequency(ies) (MHz)	Always on?	Language
ATA	India	10	No	English
BPM	China	2.5, 5, 10, 15	No	Chinese
BSF	Taiwan	5, 15	Yes	No voice
CHU	Canada	3.33, 7.335, 14.670	Yes	English/French
DUW21	Philippines	3.65	No	No voice
EBC	Spain	4.998, 15.006	No	No voice
HD2IOA	Ecuador	1.51, 3.81, 5, 7.6	No	Spanish
HLA	Korea	5	No	Korean
LOL1	Argentina	5, 10, 15	No	Spanish
LQB9	Argentina	8.167	No	No voice
LQC28	Argentina	17.551	No	No voice
PLC	Indonesia	11.440	No	No voice
PPEI	Brazil	8.721	No	No voice
PPR	Brazil	4.244, 8.634, 13.105, 17.194	No	No voice
RID	Russia	5.004, 10.004, 15.004	Yes	No voice
RTA	Russia	10, 15	No	No voice
RWM	Russia	4.996, 9.996, 14.996	Yes	No voice
ULW4	Uzbekistan	2.5, 5, 10	No	No voice
VNG	Australia	2.5, 5, 8.638, 12.984, 16	Yes	English
WWV	United States	2.5, 5, 10, 15, 20	Yes	English
WWVH	United States	2.5, 5, 10, 15	Yes	English
XBA	Mexico	6.976, 13.953	No	No voice
XDD	Mexico	13.043	No	No voice
XDP	Mexico	4.8	No	No voice
YVTO	Venezuela	5	Yes	Spanish

telephone; dial (303) 499-7111 for WWV and (808) 335-4363 for WWVH.

WWV and WWVH can be used in one of three modes.

- The audio portion of the broadcast includes seconds pulses or ticks, standard audio frequencies, and voice announcements of the UTC hour and minute. WWV uses a male voice, and WWVH uses a female voice.
- A binary time code is sent on a 100-Hz subcarrier at a rate of 1 bit per second. The time code contains the hour, minute, second, year, day of year, leap second, and Daylight Saving Time (DST) indicators and a UT1 correction. This code can be read and displayed by radio clocks.
- The carrier frequency can be used as a reference for the calibration of oscillators. This is done most often with the 5- and 10-MHz carrier signals, since they match the output frequencies of standard oscillators.

The time broadcast by WWV and WWVH will be late when it arrives at the user's location. The time offset depends upon the receiver's distance from the transmitter but should be <15 msec in the continental United States. A good estimate of the time offset requires knowledge of HF radio propagation. Most users receive a signal that has traveled up to the ionosphere and was reflected back to earth. Since the height of the ionosphere changes, the path delay also changes. Path delay variations limit the received frequency uncertainty to parts in 10^9 when averaged for 1 day.

HF radio stations such as WWV and WWVH are useful for low-level applications, such as the synchronization of analog and digital clocks, simple frequency calibrations, and calibrations of stopwatches and timers. However, LF and satellite signals are better choices for more demanding applications.

2. LF Radio Signals (Including WWVB)

Before the advent of satellites, low-frequency (LF) signals were the method of choice for time and frequency transfer. While the use of LF signals has diminished in the laboratory, they still have a major advantage—they can be received indoors without an external antenna. This makes them ideal for many consumer electronic products that display time-of-day information.

Many time and frequency stations operate in the LF band from 30 to 300 kHz (Table IX). These stations lack

TABLE IX LF Time and Frequency Broadcast Stations

Call sign	Country	Frequency (kHz)	Always on?
DCF77	Germany	77.5	Yes
DGI	Germany	177	Yes
HBG	Switzerland	75	Yes
JG2AS	Japan	40	Yes
MSF	United Kingdom	60	Yes
RBU	Russia	66.666	No
RTZ	Russia	50	Yes
TDF	France	162	Yes
WWVB	United States	60	Yes

the bandwidth needed to provide voice announcements, but they often provide both an on-time pulse and a time code. The performance of the received signal is influenced by the path length and signal strength. Path length is important because the signal is divided into ground wave and sky wave. The ground wave signal is more stable. Since it travels the shortest path between the transmitter and the receiver, it arrives first and its path delay is much easier to estimate. The sky wave is reflected from the ionosphere and produces results similar to HF reception. Short paths make it possible to track the ground wave continuously. Longer paths produce a mixture of sky wave and ground wave. And over very long paths, only sky wave reception is possible.

Signal strength is also important. If the signal is weak, the receiver might search for a new cycle of the carrier to track. Each time the receiver adjusts its tracking point by one cycle, it introduces a phase step equal to the period of a carrier. For example, a cycle slip on a 60-kHz carrier introduces a 16.67-μsec phase step. However, a strong ground wave signal can produce very good results—a LF receiver that continuously tracks the same cycle of a ground wave signal can transfer frequency with an uncertainty of about 1×10^{-12} when averaged for 1 day.

NIST operates LF radio station WWVB from Fort Collins, Colorado, at a transmission frequency of 60 kHz. The station broadcasts 24 hr per day, with an effective radiated output power of 50 kW. The WWVB time code is synchronized with the 60-kHz carrier and contains the year, day of year, hour, minute, second, and flags that indicate the status of DST, leap years, and leap seconds. The time code is received and displayed by wristwatches, alarm clocks, wall clocks, and other consumer electronic products.

3. Global Positioning System (GPS)

The Global Positioning System (GPS) is a navigation system developed and operated by the U.S. Department of Defense (DoD) that is usable nearly anywhere on earth. The system consists of a constellation of at least 24 satellites that orbit the earth at a height of 20,200 km in six fixed planes inclined 55° from the equator. The orbital period is 11 hr 58 min, which means that a satellite will pass over the same place on earth twice per day. By processing signals received from the satellites, a GPS receiver can determine its position with an uncertainty of <10 m.

The satellites broadcast on two carrier frequencies: L1 at 1575.42 MHz and L2 at 1227.6 MHz. Each satellite broadcasts a spread spectrum waveform, called a *pseudo-random noise* (PRN) code, on L1 and L2, and each satellite is identified by the PRN code it transmits. There are two types of PRN codes. The first type is a *coarse acquisition* (C/A) code, with a chip rate of 1023 chips per millisecond. The second is a *precision* (P) code, with a chip rate of 10,230 chips per millisecond. The C/A code is broadcast on L1, and the P code is broadcast on both L1 and L2. GPS reception is line-of-sight, which means that the antenna must have a clear view of the sky.

Each satellite carries either rubidium or cesium oscillators, or a combination of both. These oscillators are steered from DoD ground stations and are referenced to the United States Naval Observatory time scale, UTC(USNO), which by agreement is always within 100 nsec of UTC(NIST). The oscillators provide the reference for both the carrier and the code broadcasts.

a. GPS one-way measurements. GPS one-way measurements provide exceptional results with only a small amount of effort. A GPS receiver can automatically compute its latitude, longitude, and altitude using position data received from the satellites. The receiver can then calibrate the radio path and synchronize its on-time pulse. In addition to the on-time pulse, many receivers provide standard frequencies such as 5 or 10 MHz by steering an OCXO or rubidium oscillator using the satellite signals. GPS receivers also produce time-of-day and date information.

A quality GPS receiver calibrated for equipment delays has a timing uncertainty of about 10 nsec relative to UTC(NIST) and a frequency uncertainty of about 1×10^{-13} when averaged for 1 day.

b. GPS common-view measurements. The *common-view* method synchronizes or compares time standards or time scales at two or more locations. Common-view GPS is the primary method used by the BIPM to collect data from laboratories that contribute to TAI.

There are two types of GPS common-view measurements. *Single-channel common-view* requires a specially designed GPS receiver that can read a tracking schedule. This schedule tells the receiver when to start making measurements and which satellite to track. Another user

at another location uses the same schedule and makes simultaneous measurements from the same satellite. The tracking schedule must be designed so that it chooses satellites visible to both users at reasonable elevation angles. *Multichannel common-view* does not use a schedule. The receiver simply records timing measurements from all satellites in view. In both cases, the individual measurements at each site are estimates of (Clock A − GPS) and (Clock B − GPS). If the data are exchanged, and the results are subtracted, the GPS clock drops out and an estimate of Clock A − Clock B remains. This technique allows time and frequency standards to be compared directly even when separated by thousands of kilometers. When averaged for 1 day, the timing uncertainty of GPS common-view is <5 nsec, and the frequency uncertainty is $<1 \times 10^{-13}$.

c. GPS carrier phase measurements. Used primarily for frequency transfer, this technique uses the GPS carrier frequency (1575.42 MHz) instead of the codes transmitted by the satellites. Carrier phase measurements can be one-way or common-view. Since the carrier frequency is more than 1000 times higher than the C/A code frequency, the potential resolution is much higher. However, taking advantage of the increased resolution requires making corrections to the measurements using orbital data and models of the ionosphere and troposphere. It also requires correcting for cycle slips that introduce phase shifts equal to multiples of the carrier period ($\cong$635 psec for <1). Once the measurements are properly processed, the frequency uncertainty of common-view carrier phase measurements is $<1 \times 10^{-14}$ when averaged for 1 day.

C. Internet and Telephone Time Signals

One common use of time transfer is to synchronize computer clocks to the correct date and time-of-day. This is usually done with a time code received through an Internet or telephone connection.

1. Internet Time Signals

Internet time servers use standard timing protocols defined in a series of RFC (Request for Comments) documents. The three most common protocols are the Time Protocol, the Daytime Protocol, and the Network Time Protocol (NTP). An Internet time server waits for timing requests sent using any of these protocols and sends a time code in the correct format when a request is received.

Client software is available for all major operating systems, and most client software is compatible with either the Daytime Protocol or the NTP. Client software that uses the Simple Network Time Protocol (SNTP) makes the same timing request as an NTP client but does less processing and provides less accuracy. Table X summarizes the various protocols and their port assignments, or the port where the server "listens" for a client request.

NIST operates an Internet time service using multiple servers distributed around the United States. A list of IP addresses for the NIST servers and sample client software can be obtained from the NIST Time and Frequency Division web site: http://www.boulder.nist.gov/timefreq. The uncertainty of Internet time signals is usually <100 msec, but results vary with different computers, operating systems, and client software.

2. Telephone Time Signals

Telephone time services allow computers with analog modems to synchronize their clocks using ordinary telephone lines. These services are useful for synchronizing computers that are not on the Internet or that reside behind an Internet firewall. One example of a telephone service is NISTs Automated Computer Time Service (ACTS), (303) 494-4774.

TABLE X Internet Time Protocols

Protocol name	Document	Format	Port assignment(s)
Time protocol	RFC-868	Unformatted 32-bit binary number contains time in UTC seconds since January 1, 1900	Port 37, tcp/ip, udp/ip
Daytime protocol	RFC-867	Exact format not specified in standard. Only requirement is that the time code is sent as ASCII characters	Port 13, tcp/ip, udp/ip
Network time protocol (NTP)	RFC-1305	The server provides a data packet with a 64-bit time stamp containing the time in UTC seconds since January 1, 1900, with a resolution of 200 psec. NTP provides an accuracy of 1 to 50 msec. The client software runs continuously and gets periodic updates from the server.	Port 123, udp/ip
Simple network time protocol (SNTP)	RFC-1769	The data packet sent by the server is the same as NTP, but the client software does less processing and provides less accuracy.	Port 123, udp/ip

ACTS requires a computer, a modem, and client software. When a computer connects to ACTS it receives a time code containing the month, day, year, hour, minute, second, leap second, and DST indicators and a UT1 correction. The last character in the ACTS time code is the on-time marker (OTM). To compensate for the path delay between NIST and the user, the server sends the OTM 45 msec early. If the client returns the OTM, the server can calibrate the path using the *loop-back* method. Each time the OTM is returned, the server measures the round-trip path delay and divides this quantity by 2 to estimate the one-way path delay. This path calibration reduces the uncertainty to <15 msec.

V. CLOSING

As noted earlier, time and frequency standards and measurements have improved by about nine orders of magnitude in the past 100 years. This rapid advancement has made many new products and technologies possible. While it is impossible to predict what the future holds, we can be certain that oscillator Q's will continue to get higher, measurement uncertainties will continue to get lower, and new technologies will continue to emerge.

SEE ALSO THE FOLLOWING ARTICLES

MICROWAVE COMMUNICATIONS • QUANTUM MECHANICS • RADIO SPECTRUM UTILIZATION • REAL-TIME SYSTEMS • SIGNAL PROCESSING • TELECOMMUNICATIONS

BIBLIOGRAPHY

Allan, D. W., Ashby, N., and Hodge, C. C. (1997). "The Science of Timekeeping," Hewlett–Packard Application Note 1289, United States.

Hackman, C., and Sullivan, D. B. (eds.) (1996). "Time and Frequency Measurement," American Association of Physics Teachers, College Park, MD.

IEEE Standards Coordinating Committee 27 (1999). "IEEE Standard Definitions of Physical Quantities for Fundamental Frequency and Time Metrology—Random Instabilities," Institute of Electrical and Electronics Engineers, New York.

ITU Radiocommunication Study Group 7 (1997). "Selection and Use of Precise Frequency and Time Systems," International Telecommunications Union, Geneva, Switzerland.

Jespersen, J., and Fitz-Randolph, J. (1999). "From Sundials to Atomic Clocks: Understanding Time and Frequency," NIST Monograph 155, U.S. Government Printing Office, Washington, DC.

Kamas, G., and Lombardi, M. A. (1990). "Time and Frequency Users Manual," NIST Special Publication 559, U.S. Government Printing Office, Washington, DC.

Levine, J. (1999). "Introduction to time and frequency metrology," *Rev. Sci. Instrum.* **70,** 2567–2596.

Seidelmann, P. K. (ed.) (1992). "Explanatory Supplement to the Astronomical Almanac," University Science Books, Mill Valley, CA.

Sullivan, D. B., Allan, D. W., Howe, D. A., and Walls, F. L. (eds.) (1990). "Characterization of Clocks and Oscillators," NIST Technical Note 1337, U.S. Government Printing Office, Washington, DC.

Vig, J. R. (1992). "Introduction to Quartz Frequency Standards," United States Army Research and Development Technical Report SLCET-TR-92-1.

Walls, F. L., and Ferre-Pikal, E. S. (1999). Frequency standards, characterization. Measurement of frequency, phase noise, and amplitude noise. *In* "Wiley Encyclopedia of Electrical and Electronics Engineering," Vol. 7, pp. 767–774, Vol. 12, pp. 459–473, John Wiley and Sons, New York.

Ultrasonics and Acoustics

David R. Andrews
Cambridge Ultrasonics

GLOSSARY

Anti-node A point of maximum amplitude of vibration (see node).
Array A collection of transducers generally arranged with constant separation.
Bandwidth A range or band of frequencies.
Chirp A burst of sinusoidal waves.
Deconvolution Reverse of the process of convolution. An example of convolution is the response of a mechanical system to a stimulus.
Group velocity The speed at which energy and information is carried by a wave.
Harmonics If a vibration has a frequency f then its even harmonics will be at $2f$, $4f$, $6f$, ... the odd harmonics will be at $3f$, $5f$, $7f$... and the *subharmonics* will be at $f/2$, $f/3$, $f/4$, ...
Mode-conversion The process by which compression strain is converted into shear strain at a surface or vice versa.
Node A point of low or zero amplitude of vibration.
Phase velocity The speed at which a single frequency component travels.
Pitch Term used in music for frequency.
Q-factor Quality of resonance. Center frequency divided by -3 dB bandwidth or the time for oscillations to dampen to $1/e$ divided by the period of the oscillation.
Shear A transverse relative motion of particles.
White noise Random noise covering a wide (infinite) range of frequencies.

MECHANICAL vibrations and waves in solids, liquids, and gases can be classed as ultrasonic or acoustic waves.

Liquids and gases only support compression waves, represented by a scalar pressure, with a longitudinal particle motion in the same direction as the wave. In solids there can also be two orthogonal, transverse motions (vector shear strains), all of which can be combined into a single tensor representation. Ultrasonic and acoustic waves offer considerable scope for technological exploitation because of the diversity of wave modes, the ease with which waves can be launched and received, their ability to travel long distances, and because the waves are inherently sensitive to mechanical structure. They have found application in many fields, most notably: medical diagnosis, nondestructive evaluation, geophysical exploration, and sonar. Applications using pulsed waves are predominantly associated with inspection whereas applications using continuous waves are predominantly associated with material processing. The spectrum of acoustics and ultrasound covers three ranges of frequencies, the first from about 0.1 to 20 Hz is sometimes termed *infrasound*; the second is the range of human hearing, generally taken to be 20 Hz–20 kHz; the third extends beyond human hearing up to about 100 MHz. It is this third range that is commonly termed *ultrasound*. Use of the term *acoustic* lacks precision for it is sometimes used to indicate the range of human hearing while at other times it is used to include infrasound and ultrasound as well. There are relatively few technological applications in the range of human hearing, apart from musical instruments and systems used to relay audible information to humans, the reason being, presumably, to avoid causing interference to human hearing. Experiments show that the attenuation of ultrasonic waves increases with increasing frequency for virtually all materials. At a frequency of 100 MHz the wavelength of ultrasound is so small, typically 50 μm in metals, 15 μm in liquids such as water, and 3 μm in gases, that all materials have significantly high attenuation and 100 MHz is the approximate upper limit of technological applications of ultrasound.

I. FUNDAMENTALS

Any system, comprising a collection of masses, that is able to apply forces between the masses can carry ultrasonic or acoustic waves. Distributed masses or point masses, quasi-static forces that obey Hooke's law (termed Hookean forces), or very short duration collision forces can all support acoustic waves. More specific examples of such systems are, at the microscopic scale, atoms and molecules in solids, liquids, and gases; at a very large scale galaxies should be capable of supporting very low frequency waves.

A. One-Dimensional System

The simplest system able to support ultrasonic or acoustic waves has only one dimension, such as a rope or string held taught at each end, where mass is continuously distributed along the length and the force is the line tension. A one-dimensional system can support the following two fundamental modes of vibration.

1. Longitudinal or compression waves (scalar). The compression of the wave at any point along the string can be described by a scalar quantity. Particle motion is parallel to the direction of travel of the wave.
2. Transverse waves (vector). The motion of particles in a transverse wave is perpendicular to the direction of travel of the wave. The transverse displacement is described by resolving it into two orthogonal planes. It is possible to have polarization states of transverse waves, in which two orthogonal waves of the same frequency and speed have a fixed phase relationship, for example, linear, circular and elliptical polarizations.

Real ropes and strings have a measurable thickness and can also support torsion vibrations, due to the moment of inertia and the shear modulus of the string. Compression (longitudinal scalar) and shear (transverse vector) are the two fundamental forces and motions in ultrasonic and acoustic waves.

B. More Complex Systems

Periodicity in a system causes periodicity in the vibration pattern and any solution must satisfy Floquet's principle.

$$F(z + d) = F(z)$$

Where $F(z)$ describes the vibration pattern and d is the periodicity. A Fourier series of the following type is a solution because of Floquet's principle.

$$F(z) = \sum_{n=\infty}^{\infty} a_n e^{-i(2\pi n/d)z}$$

The jointed pipes used in a riser in an oil well are an example of a periodic structure. Floquet's principle shows the riser behaves like a filter to acoustic waves with periodic pass-bands and nulls (comb-filter).

In the one-dimensional systems considered so far the masses were distributed evenly and the forces obeyed Hooke's law. Force is proportional to extension in Hooke's law. Newton's law, relating force, mass, and acceleration are used with Hooke's law to construct an equation of motion. The same principles are applied in three dimensions but tensor notation is used. The concept of a force is

TABLE I Data for Five Gases of Different Molecular Weights Showing the Variation of Sound Speed with Molecular Weight. Lighter Molecules Transport Sound Faster than Heavier Molecules

Gas (at 273 K and 10^5 Pa)	Molecular weight	Speed of sound (ms^{-1})
Air	14	330
Carbon dioxide	44	260
Deuterium	2	890
Hydrogen	1	1300
Hydrogen bromide	81	200

replaced by stress, the force per unit area, and the concept of extension is replaced by strain, the extension per unit length. Both stress and strain are tensor quantities and each contains components describing compression and shear.

C. Atomic Models of Wave Transport Mechanisms—Speed of Sound

Continuum models predict that the speed of sound, c, in a gas at moderate pressures is given by

$$c = \sqrt{\gamma RT/M}$$

Where γ is the ratio of specific heats at constant pressure and constant volume, T is the temperature in Kelvin, and M is the molecular weight. This expression shows that the larger the mass of the molecule the more difficult it is to move it quickly and the lower will be the speed of sound. The factor γRT is a constant to a fair approximation.

It is known that the speed of sound, c, in a liquid is given by

$$c = \sqrt{1/\beta_a \rho}$$

Where β_a is the adiabatic compressibility and ρ is the density. Note the $1/\sqrt{\rho}$ dependency, which is the same kind of dependency as $1/\sqrt{M}$ for gases. In this case, β_a is not a constant, it is a material parameter with significant variability.

TABLE II Data for Five Liquids of Different Densities Showing the Variation of Sound Speed with Density

Liquid (at 293 K and 10^5 Pa)	Density ($kg\ m^{-3}$)	Speed of sound (ms^{-1})
Ethyl alchol	789	1100
Helium (4.2 K)	120	183
Mercury	13590	1450
Sodium (383 K)	970	2500
Water	1000	1500

TABLE III Data for Six Solids Showing the Variation of Sound Speeds for Three Different Wave Modes: Longitudinal, Transverse, and Surface Waves

Solid (at 293 K and 10^5 Pa)	Density ($kg\ m^{-3}$)	Speed of sound (longitudinal) (ms^{-1})	Speed of sound (transverse) (ms^{-1})	Speed of sound (surface) (ms^{-1})
Aluminum	2700	6400	3100	2900
Diamond	2300	18600		
Concrete	2400	2500–5000	1200–2500	1000–2000
Sapphire (Al_2O_3) z axis	4000	11000	6000	
Steel—mild steel	7900	6000	3200	3000
Wood	650	3500		

It is known that the longitudinal speed of sound, c_L, and the transverse speed of sound, c_T, in a solid are given by

$$c_L = \sqrt{E/\rho} \quad \text{and} \quad c_T = \sqrt{G/\rho}$$

Where E is Young's modulus, G is the shear modulus, and ρ is the density. The last two expressions only apply to isotropic materials. It is always true that $c_L > c_T$. A surface wave travels along a surface of a material. The particle motion is elliptical with the greatest amplitude at the surface, decaying with depth. Surface waves travel at speeds approximately the same as shear waves.

D. Wave Equation

For all four equations predicting the speed of sound in Section I.C there is a general form of

$$Speed = \sqrt{\frac{stiffness}{density}}$$

This is a consequence of the wave equation for ultrasonic and acoustic waves. In one-dimension the general form is

$$\frac{\partial^2 u}{\partial x^2} = \frac{\rho}{E}\frac{\partial^2 u}{\partial t^2}$$

General solutions (d'Alembert's solution) predict two waves traveling at speed c in both the positive x-direction and the negative x-direction.

$$u = u(x \pm ct)$$

When either solution is used then the following relationship emerges

$$1 = \frac{\rho}{E}c^2 \qquad \text{or} \qquad c = \sqrt{\frac{E}{\rho}}$$

Explicit solutions have been found for many systems with rectangular, spherical, and cylindrical geometry. Where an explicit solution cannot be found then computational methods such as the finite element and the finite difference methods can be used. They represent the vibrating system by point masses and springs located on a mesh spanning the volume of the system. Computers can be used to calculate approximate solutions.

E. Reflection, Transmission, Refraction, and Mode Conversion

When a wave in one material passes through an interface into a second material some energy is reflected and some is transmitted. By considering the continuity of the amplitudes of waves normal and parallel to the interface it can be shown that the reflected intensity is a proportion R of the incident intensity

$$R = \left(\frac{Z_2 - Z_1}{Z_2 + Z_1}\right)^2$$

Where Z_1 is the acoustic impedance of the first material and Z_2 is the acoustic impedance of the second material and acoustic impedance equals the product of density and wave speed. The proportion of the beam transmitted, T, is given by

$$T = 1 - R$$

If the angle of incidence is oblique and the angle with the normal is θ then the reflected beam is at an angle ϕ to the normal and

$$\theta = \phi$$

Snell's law applies to the refraction of ultrasonic and acoustic waves,

$$\frac{\sin\theta}{\sin\theta_2} = \frac{c_1}{c_2}$$

with θ_2 the angle of refraction to the interface, c_1 the speed of sound in the material supporting the incident wave and c_2 the speed of sound in the material supporting the refracted wave.

The effects mentioned above are well known wave effects in physics but mode conversion, as illustrated in Fig. 1, is unique to waves traveling in solids. In this instance a compression wave is converted partially into a shear wave because stresses must be zero on the surface of the solid.

F. Transducer Beams

When a wave is launched from a transducer it is interesting to know how it travels thereafter. First, plane waves emerge, with the same width as the transducer and, secondly, waves are created at its perimeter, known as edge

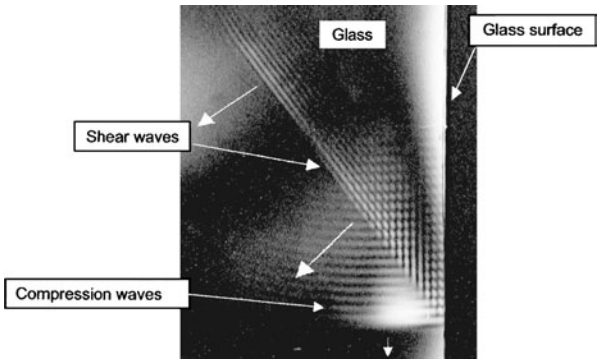


FIGURE 1 A photograph of ultrasonic waves rendered visible in glass. Compression waves at 2.5 MHz are traveling downward, grazing a free surface where mode-conversion creates shear waves at the same frequency. Since the speed of the shear waves is lower, the wavelength is shorter and they are emitted at an angle to the compression waves.

waves. If the transducer is circular then the edge wave is an expanding toroid, this wave can be seen in cross section in Figs. 2 and 3. Close to the transducer, within a distance L, in the *near field*, the edge waves interfere with each other and with the plane waves, resulting in rapidly varying amplitudes with distance. Further away, in the *far field*, the edge waves are always approximately tangential to the plane waves in the center of the beam and only constructive interference can occur.

Interference in the center of the far field is always constructive if the edge waves are within half the wavelength $\lambda/2$ of the plane wave. This criterion can be used to predict the range of the near field, L, using simple geometry. If the aperture of the transducer is D then

$$\frac{\lambda}{2} < L - \sqrt{L^2 - \left(\frac{D}{2}\right)^2}$$

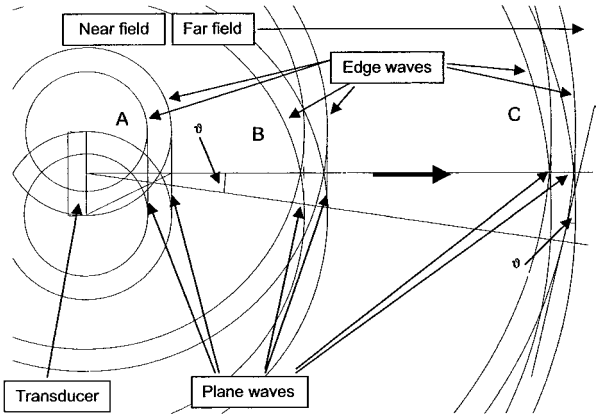


FIGURE 2 Sketch illustrating how waves emerge (traveling from left to right) from a transducer and cross three regions: A near field, B, and C the far field.

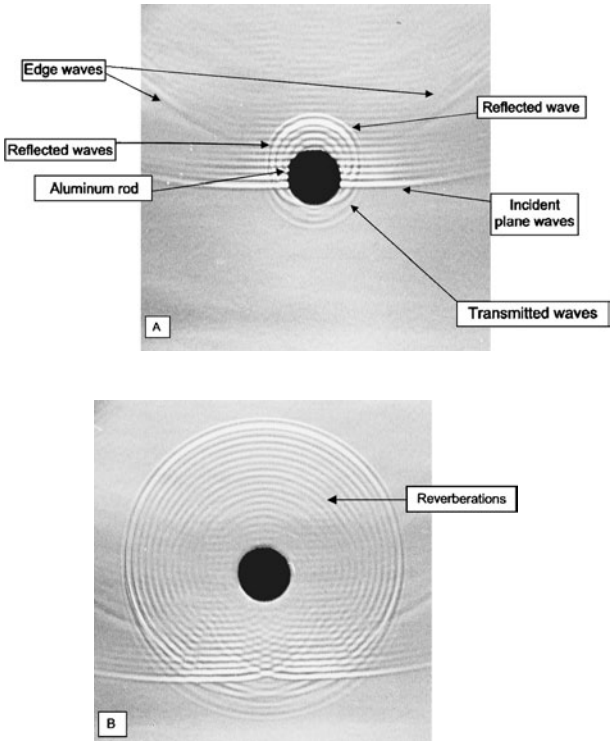


FIGURE 3 Two photographs showing ultrasonic plane waves and edge waves (1.5 MHz) rendered visible in water. Photograph B shows waves approximately 10 μs after the first. The solid, black circle is a solid aluminum cylinder or rod viewed along its length. The photographs show the effects of reflection, transmission, and reverberation.

Or to a first approximation

$$L < \frac{D^2}{4\lambda}$$

At the side of the beam in the far field, the edge waves from opposite sides of the aperture interfere destructively creating an amplitude null, or node, in the shape of a cone of semi-angle ϑ. Outside this cone the interference is successively constructive and destructive. Simple geometry can be used to estimate ϑ.

$$\sin\vartheta \approx \frac{\lambda/2}{D/2} = \frac{\lambda}{D}$$

More rigorous arguments show that

$$\sin\vartheta = 1.22\frac{\lambda}{D}$$

G. Attenuation

A mechanical wave can lose energy by two principal mechanisms: thermoelastic losses in homogeneous materials (energy is converted into heat) or scattering in heterogeneous materials (the wave is scattered in many directions). Both mechanisms result in a gradual loss of intensity as the wave travels and the effects are referred to as attenuation. Scattering does not convert ultrasonic wave energy into another form of energy so no energy is lost by this mechanism, but the wavefront loses coherence. The value of attenuation is proportional to frequency squared for thermoelastic losses. The way attenuation is measured can strongly influence the value when scattering is the dominant mechanism. The randomly distributed aggregate particles in concrete, for example, are random scatterers if the wavelength is equal to or less than the size of the aggregates. If attenuation is measured using a pair of 50-mm diameter transducers at 200 kHz (wavelength is approximately 20 mm) then the receiver will register a small signal, indicating high attenuation. The same experiment performed on a homogeneous material, like aluminum, results in a much larger signal. Concrete has a higher attenuation than aluminum, measured this way. However, when an array of 10-mm diameter receivers is used as an energy detector then a different, much lower value of attenuation is measured for concrete (see Fig. 4). Random scattering is referred to as speckle in medical ultrasonic systems.

Attenuation, α, is formally measured in nepers m^{-1} but it is more commonly measured in dB m^{-1} or sometimes

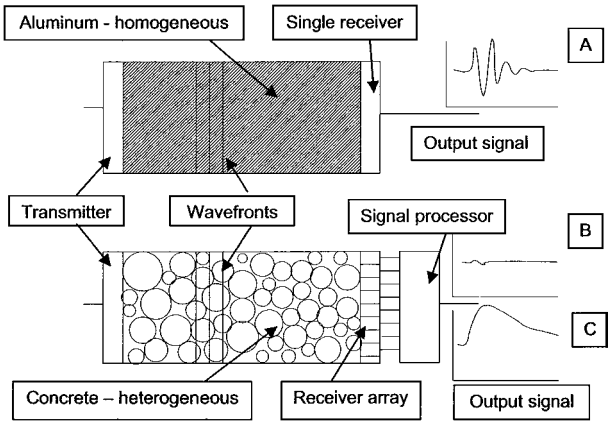


FIGURE 4 Sketch to illustrate the effect of scattering losses in an attenuation measurement and the effect of receiver type on the result. Short, identical bursts of waves are emitted by identical transmitters into samples of homogeneous and heterogeneous materials, with equal sample sizes. A strong signal (A) is collected using a single, coherent receiver from the homogeneous material but a heterogeneous material gives a small signal (B) with a coherent receiver, indicating high attenuation. However, the heterogeneous material can still give a strong signal (C) when an energy detector is used instead of a single coherent detector, indicating low attenuation. Signal processor in B—simple summing circuit. Signal processor in C—envelope detector followed by summation.

TABLE IV Variation of Attenuation for Four Materials

	Attenuation α (db m^{-1})	Attenuation/ wavelength (db)	Attenuation α (neper m^{-1})
Water 1 MHz	0.22	1.5×10^2	2.5×10^{-2}
Water 1 GHz	2.2×10^5	1.5×10^{11}	2.5×10^4
Blood 1 MHz	18	1.2×10^4	2.1
Air 1 MHz	1200	3.6×10^6	138
Aluminium 1 GHz	7500	1.3×10^9	860

in dB m^{-1} Hz^{-1}. The neper is a consequence of the exponential decay of intensity with distance.

$$\alpha(\text{db m}^{-1}) = 20(\log_{10} \text{e})\,(\text{neper m}^{-1})$$

Where e (value 2.718) is the base of natural logarithms.

It is also common to compare materials in terms of attenuation per wavelength (measured in dB) because the size of an experiment or ultrasonic system is usually an important underlying factor to consider and, generally, the size of the experiment in proportion to the wavelength. Table IV gives some typical attenuation values.

H. Doppler Effect

When ultrasound of frequency f is reflected from a scatterer, which is moving relative to the material supporting the ultrasonic wave, then the frequency of the reflected wave is changed. The amount the frequency is changed, f_d, is known as the Doppler-shift frequency and the value depends upon the vector velocity, $\underline{v}$, of the scatterer relative to the vector of the ultrasound frequency, $\underline{f}$, in the direction of travel of the ultrasonic wave (speed c). Where $\underline{v}$ is positive if it is measured in the same direction as $\underline{f}$.

$$f_d = -\frac{2}{c}\underline{v} \cdot \underline{f}$$

The Doppler effect is exploited in medical ultrasonic systems to measure the speed of flowing blood or the speed of heart valves opening and closing. It is also used in some flow meters for metering fluids, for example, water or oil and gas. The doppler shift frequency for a wave of 2 MHz passing through blood flowing at 10 cm s^{-1} at 60° away from the transducer is −125 Hz (frequency reduction). The frequency of the returning echo will be 1,999,875 Hz. In compound B-mode images color is used to indicate the presence of doppler shifts.

I. Dispersion

Experiments show that materials generally cause pulses of waves to become longer as they travel. This effect is called dispersion. Dispersion is caused by waves of different frequencies traveling at different speeds, with the lowest frequencies usually traveling fastest. Dispersion and attenuation are closely related. There are many causes of dispersion including: the existence of boundaries to the material, particularly if they are regular or symmetrical; inherent material properties at the molecular and atomic level, associated with force transfer and effective mass; scattering of waves in heterogeneous materials; and the dependence of wave speed on amplitude (nonlinear dispersion).

A mathematical expression for a traveling wave is

$$\sin(kx - \omega t) + \sin(kx + \omega t)$$

Where x is the distance traveled, t is time, $k = 2\pi/\lambda$ is the wave number (λ is the wavelength) and $\omega = 2\pi f$ is the angular frequency (f is frequency). The phase velocity or the speed of a single frequency component in the wave is $v = f\lambda = \omega/k$. The speed at which energy or modulation moves along with the wave, v_g, is called the group velocity. It is given by

$$v_g = \frac{d\omega}{dk} = v + k\frac{dv}{dk}$$

The effect of dispersion on a pulse is progressive, depending upon the distance traveled in the material. The study of dispersion can be aided by using time-frequency representations of transmitted signals.

J. Sonar Equation

The sonar equation is of fundamental importance in all ultrasonic and acoustic systems although it is more widely used in sonar design than in other applications. It is used to predict the voltage level, R_x, of an echo from a target given an electrical voltage drive, T_x, to the transmitter. In simplified form it is

$$R_x = T_x \cdot C_x^2 \cdot \frac{S}{r^4}$$

Where C_x is the electromechanical conversion efficiency of the transducer, assumed here to be used both for transmission and reception hence raised to a power of two, S is the strength of the scatterer (how much incident energy it reflects back), and r is the range to the scatterer, assuming an omnidirectional transducer in a deep ocean. The power of the transmitter is spread out over a spherically expanding shell of area $4\pi r^2$ as it travels to the scatterer. At the scatterer some power is reflected back and it becomes a secondary source, creating a second spherically expanding shell of area $4\pi r^2$, which travels back to the transmitter/receiver. It is the combined effect of the two expanding shells that accounts for the $1/r^4$ term. Factors can also be included to allow for attenuation during transmission.

TABLE V Quantities Commonly Used in the Field of Ultrasonics and Acoustics

Quantity	Unit name	Unit symbol	Derivation
Attenuation			neper m^{-1} (or dB m^{-1})
Density			kg m^{-3}
Frequency	Hertz	Hz	s^{-1}
Impedance (acoustic)	Rayl	Z	kg m^{-2} s^{-1}
Intensity			W m^{-2}
Pressure	Pascal	Pa	N m^{-2}
Speed			m s^{-1}
Wavelength			m

K. Derived SI Units Encountered in Ultrasonics and Acoustics

Table V lists the majority of units used in the field of ultrasonics and acoustics. Impedance = density × speed and intensity = energy/area.

L. Common Transducers

Below 100 kHz electromagnetic and electrostatic devices can be used as the active electromechanical components in transducers. Up to about 50-kHz magnetostrictive devices are popular for power transmitters. A magnetostrictive material deforms and generates strain energy when a magnetic field is applied to it. Above about 100 kHz it is common to use ferroelectric ceramics such as lead zirconate titanate (PZT) as the electromechanical devices in transducers. Ferroelectric materials can be electrically poled to impart apparent piezoelectric properties to them. A piezoelectric material deforms and creates strain energy when an electric field is applied to it. Piezoelectric ceramics have relatively high Q or quality factors but one exception is modified lead metaniobate, which is naturally damped.

A commonly used shape for a PZT component is a thin disc, it has three principal modes of vibration: flexural, generally of low frequency, through the thickness and along the radius. If the thickness of the disc is less than its radius then the frequency of the thickness mode will be the highest of all the modes, and simple electrical circuits can be used to suppress the low frequency modes. Discs of PZT can be used in this way to make transducers with relatively simple frequency responses based about chosen center frequencies. Manufacturers usually describe transducers in terms of the center frequency, aperture size, and wave type (compression or shear). Sometimes the number of cycles resulting from a single impulse excitation is also given, for example, two and a half cycles ring-down, which is related

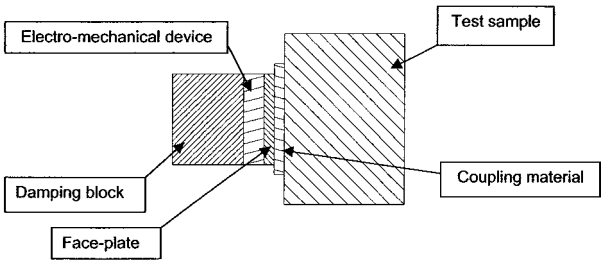


FIGURE 5 Sketch to illustrate the construction of an ultrasonic transducer with a working frequency greater than 100 kHz.

to the frequency response and the Q of the transducer. Although shear-wave transducers can be made from PZT in the form of shear-plates it is more common to use compression discs, angling the disc inside the transducer body at the transmitting surface at, for example, 45° instead of 90° then shear waves are generated by mode-conversion at the interface with the test sample and the compression waves in the transducer are reflected internally and absorbed.

Commercial transducers (see Fig. 5) generally have a block of tungsten-loaded epoxy resin bonded to the surface of the PZT disc that is not coupled to the sample. The block is typically several wavelengths long and provides strong mechanical damping. A transducer also has a face-plate to protect the disc, generally in the form of a thin polymer membrane. In some instances the face-plate is made thicker, from an epoxy-resin material, to be a quarter-wave plate. The thickness is made equal to a quarter of the wavelength at the center frequency of the transducer and the acoustic impedance is made equal to the geometric mean of the piezoelectric disc and the test sample. Under these special circumstances the face-plate acts as a matching layer between the disc and the material under test, maximizing the coupling of energy between the test sample and transducer. The transducer must be mechanically coupled to the test sample while testing. Common commercial coupling materials are water-based gels but water, grease, sodium salicylate, and adhesives can be used. It is possible to use air as the coupling material, but this results in exceptionally high signal losses because of the high attenuation of air and very low transmission coefficients between transducer materials and air (acoustic impedance mismatch).

Electromagnetic and electrostatic devices are popular electromechanical elements in microphones and loudspeakers. A moving coil loudspeaker has a coil of wire suspended in the strongest part of the field of a permanent magnet (see Fig. 6). A lightweight, rigid cone, supports the coil at its apex and is loosely supported in turn at its outer perimeter by a rigid metal frame. The cone is free to move through a distance of several millimeters parallel to the axis of the coil.

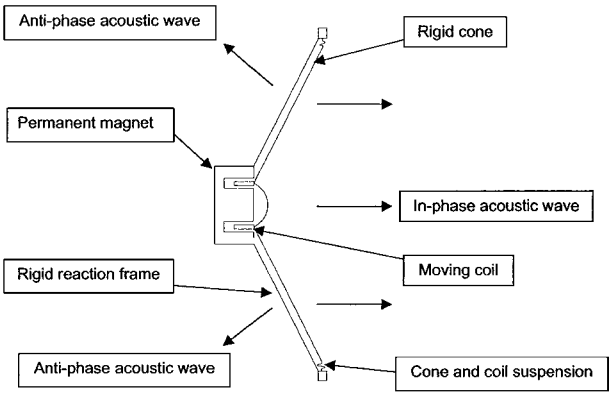


FIGURE 6 Sketch showing a cross section through a loudspeaker drive unit.

When an electric current passes though the coil a force is generated parallel to the axis that causes the coil and cone to move, compressing a large volume of air and creating a sound wave. Without the cone the coil would be virtually inaudible. When the coil moves out from the magnet it compresses the air moving toward the listener, but on the opposite side of the cone the air is rarefied. The rear of the loudspeaker therefore generates an anti-phase acoustic signal. It is common to include two or three loudspeaker drive units of different sizes into a single assembly to improve the overall fidelity of reproduction. The smallest drive unit transmits the high frequency sounds (usually above 1 kHz) and the largest drive unit transmits the low frequency sounds (usually below 200 Hz). Electrical filters, made of capacitors and inductors, known as cross-overs, direct electrical energy of an appropriate frequency range to each drive unit. The drive units are mounted in one housing, which contributes to the tonal quality of the final sound. The housing is sealed, apart from one aperture, which provides a path for the release of pressure from the inside of the housing and it is packed with material to absorb anti-phase acoustic energy. This construction is called an *infinite baffle* because the loudspeakers operate as if the rear of the cones were coupled to a semi-infinite volume of air, which presents the minimum mechanical load. The anti-phase sounds generated inside the assembly should not be transmitted to the listener because they would interfere with the in-phase acoustic waves and degrade the fidelity of reproduction. The working frequency range of moving-coil loudspeakers is approximately 20 Hz–20 kHz.

An electrostatic force is created when two exposed electric charges are brought close together. In an electrostatic transducer one of the electric charges is distributed over a thin, metalized polymer film which is separated by a distance of a few millimeters from the second electric charge on a rigid metal plate. The film and plate form an

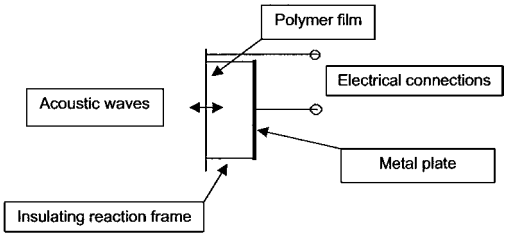


FIGURE 7 Sketch of an electrostatic transducer. The two electrical terminals are connected to a polarizing voltage and also to either a source of modulation (case of a transmitter) or to a high impedance amplifier (case of a microphone).

air-filled capacitor with the film free to vibrate. A generator of electrostatic charge maintains approximately constant charges on the film and plate. The device acts as a microphone if an acoustic wave reaches the polymer film and the resulting charge modulations are used as a signal or it can be used as a loudspeaker if the charges are modulated by a driving signal. Electrostatic microphones are intrinsically simple and can be made as small as 10 mm in size. Electrostatic loudspeakers are generally made with much larger transmitting areas (up to 1 m^2) to give higher efficiency. The working frequency range of electrostatic transducers is approximately 20 Hz–100 kHz.

II. PULSED METHODS

Pulsed ultrasound is used for inspection purposes including: medical imaging, sonar, acoustic microscopy, and nondestructive evaluation. In a pulsed system a transducer is placed in contact with a sample, electrical driving signals are sent to the transducer to make pulses of ultrasonic waves, which enter the sample and echoes from internal scatterers are collected by a receiver, processed, and displayed for interpretation (see Fig. 8).

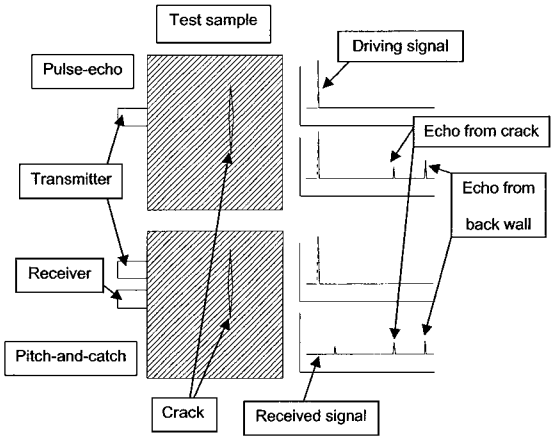


FIGURE 8 Conventional pulsed methods used for inspecting a test sample, showing impulse drive signals and received signals.

If the transmitter is used as a receiver the method is known as pulse-echo, but if a second transducer is used as a receiver then the method is known as pitch-and-catch. The times of arrival of echoes can be measured and converted into distance using the appropriate value for the speed of sound in the sample and other factors depending on the geometry of the test (division by two in this case). As well as having echoes from cracks in the sample the received signal may also have echoes from surfaces, such as the back-wall echo shown in Fig. 8, to add to the difficulty of interpretation. Further complications arise with parallel-sided, objects with high Q factors that reverberate, generating relatively long series of exponentially damped pulses (see Fig. 3B).

A. Axial and Lateral Resolution

The value of the axial resolution is equal to the spatial extent of the pulse in the test material and is the pulse duration multiplied by the wave speed. Two scatterers can be resolved (axially) as distinct along the acoustic axis if their echoes are separated by more than the axial resolution, therefore, a short ultrasonic pulse gives better resolution than a long pulse. Transducers with higher center frequencies emit shorter pulses and are used when finer axial resolution is required. Two scatterers can be resolved (laterally) as distinct in a direction perpendicular to the acoustic axis if their echoes are separated by more than the beam width. Lateral resolution is determined mainly by the transducer aperture size for collimated transducers. A diverging beam has a lateral resolution that increases with range and a focused beam has the smallest lateral resolution in the vicinity of the focus—after the focus the beam diverges. It is generally found that focusing transducers give the best resolution but only in the vicinity of the focus.

B. Driving Signals and Processing

A chirp signal is a burst of waves that can be synthesized (referred to here as controlled chirps) or otherwise created and processed. However, chirps are frequently used inadvertently in pulsed ultrasonic and acoustic systems because all transducers respond with a chirp of some description when driven electrically, irrespective of the drive signal. Many systems do little to process the chirp automatically. The most common and least effective signal processing of chirps is human interpretation of a displayed signal; the alternative is automatic processing by computer. Matched filtering, which is a form of linear pattern detection, converts a long chirp into a short impulse (compression) and can only be done quickly by computer. Compression improves the axial resolution and this is the main benefit of matched filtering. Matched filtering is closely connected with deconvolution (see Section VIII).

Figure 9 follows two chirps, a controlled chirp and an *inadvertent* chirp (impulse drive), through typical stages of processing. Two signals were synthesized by convoluting the transducer's response with each of the two electrical drive signals, and convoluting the result with a perfect echo at 1000 time samples and adding white noise. At this stage signals represent typical received signals. The next stage processed the inadvertent chirp in two ways, the first being typical of many conventional, nondestructive test systems was rectification followed by level detection; the second way was applying deconvolution, namely, matched filtering followed by envelope detection. The controlled chirp was only processed by deconvolution. The match filter was the time-reversed recording of the output of the transducer. Figure 9C shows a typical received signal. Figure 9D shows the signal presented for interpretation by a conventional nondestructive ultrasonic test showing three peaks from one echo, the peaks come from the inadvertently formed chirp (due to the impulse response of the transducer). Figure 9E shows the conventional test with deconvolution, there is now only one main peak instead of three. Figure 9F shows deconvolution of the controlled chirp with only one peak and low background noise.

In summary, many commercial, pulsed inspection systems are used successfully with inadvertently formed chirps, but there are some advantages to be gained in using controlled chirps and deconvolution when axial resolution is important.

C. Multiplicative and Additive Noise

Multiplicative noise is due to random scattering in the material under test and it is coherent with the driving signal to a varying degree. Averaging of signals collected at several locations (spatial averaging) within a few wavelength's distance is effective in improving the detection of extended targets in the presence of multiplicative noise. Averaging of signals collected at different times (time averaging) is effective in improving the signal-to-noise when additive, noncoherent noise is present. The improvement in additive signal-to-noise is proportional to $\sqrt{N}$, where N is the number of signals averaged.

D. Common Ways of Using Pulsed-Ultrasound

Some common classes of pulsed inspection methods are known as: A-scans (amplitude or A-mode), B-scans (brightness B-mode), compound B-mode scans, and C-scans (two-dimensional scanning). The terminology is virtually the same in medical scanning and nondestructive

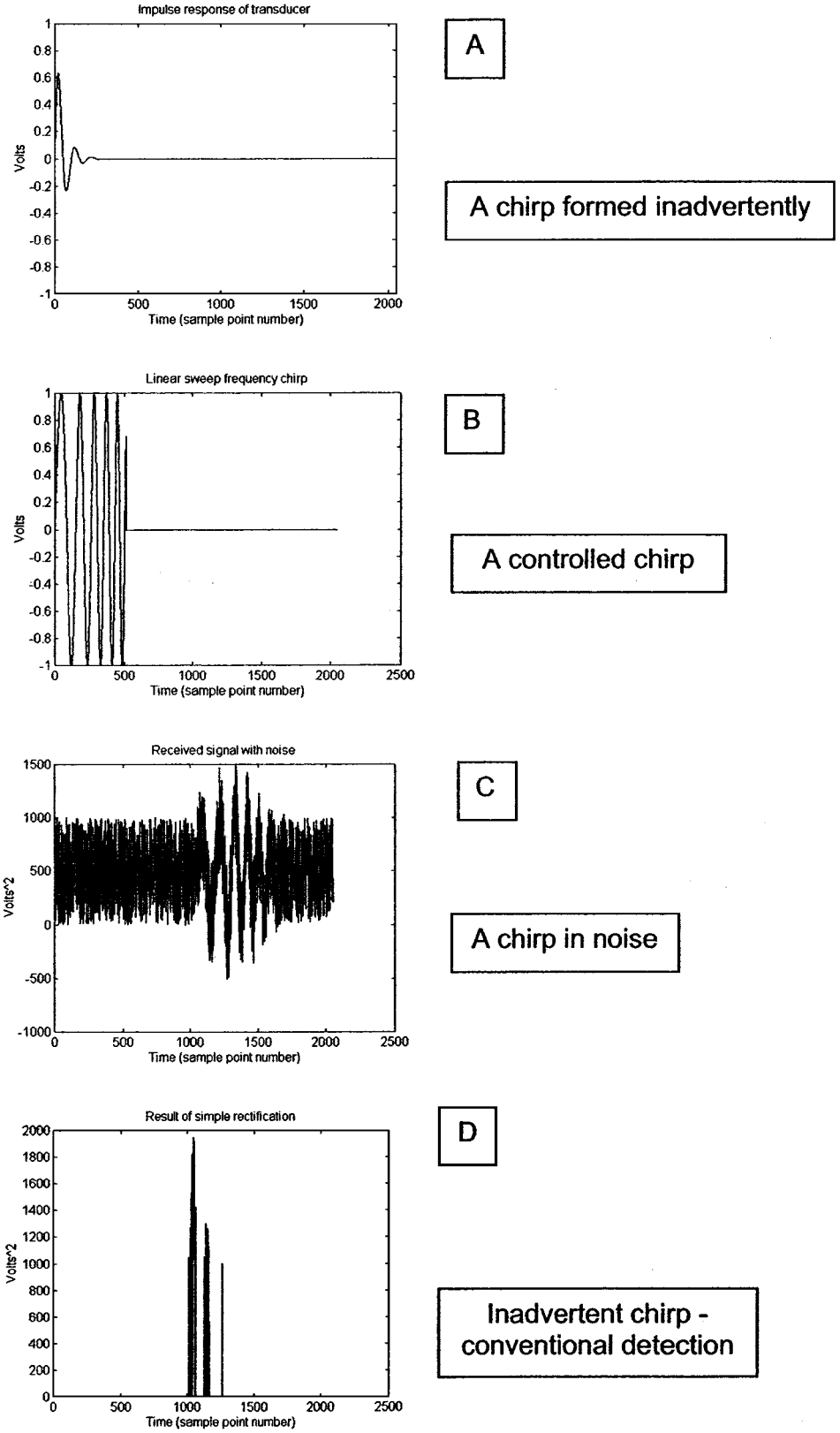


FIGURE 9 Comparison of the signals used in a conventional, pulsed inspection system (inadvertently formed chirps) and a system using controlled chirps with deconvolution.

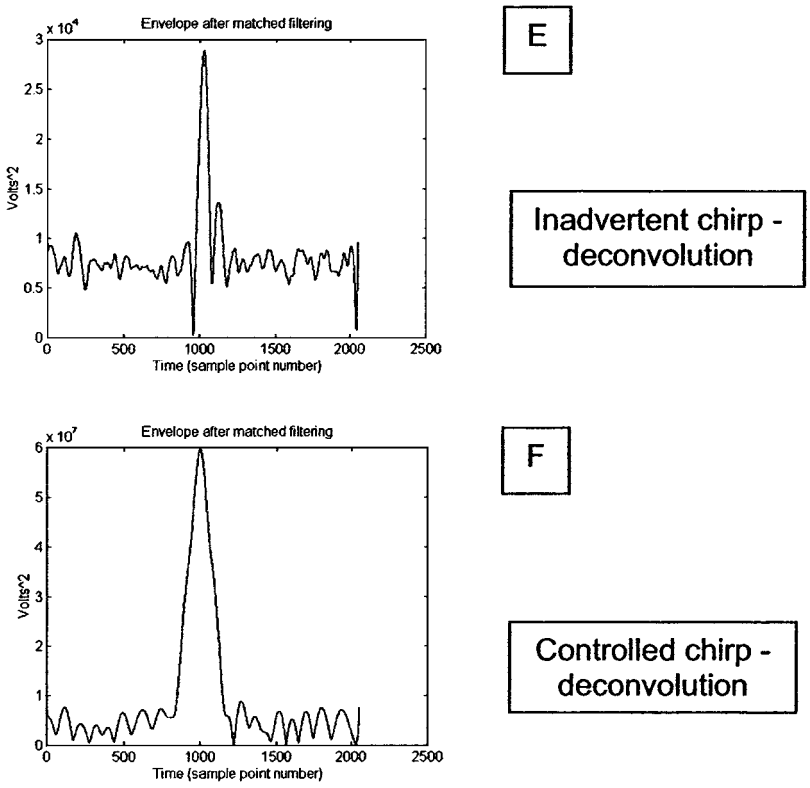


FIGURE 9 (*Continued*)

evaluation. Pulse-echo is commonly used in all three scanning methods. In an A-mode scan the received signal is displayed using two dimensions with time as the x axis and amplitude the y axis (see Fig. 9). B-mode is similar to A-mode but there is no y axis deflection on the display, just a straight line parallel to the x axis whose brightness at any time is controlled by the amplitude of the ultrasonic signal. A compound B-mode scanned image is built-up of many B-mode scans, with the transducer scanning the test sample. Each B-mode line on the display is displaced from the others in such a way as to represent the physical scanning of the beam. The compound B-mode scan provides a two-dimensional image, commonly slicing through the sample. C-scanning also creates a two-dimensional image but one in which the image-plane is perpendicular to the acoustic axis, requiring two-dimensional scanning by the transducer of the sample. There are two popular embodiments of C-scanning, both use fluid coupling between the transducer and test sample: the acoustic microscope (scanning typically 1×1 mm area) and the water immersion tank (scanning typically 100×100 mm area). The latter can take several minutes to form an image. A transducer is generally focused on the test sample in a C-scan and the amplitude of the focused portion of the received signals controls the brightness of the display for each point in the scan.

III. CONTINUOUS WAVE METHODS

A single continuous wave (CW) is, in theory, a wave of infinite duration and is inherently incapable of providing axial resolution for imaging purposes. For this reason CW is never used for imaging and is seldom used for non-destructive evaluation. However, CW is better suited to high power applications than pulsed ultrasound. Examples of typical CW applications are ultrasonic cleaning baths, agitators for processing materials, and medical therapy instruments.

A. Sources of High Power CW Ultrasound

A source of CW ultrasound is generally a resonating structure with a high quality factor. It is pumped with mechanical energy by an active electromechanical device at its resonating frequency. In high-power applications the design objective is to maximize the amplitude of vibration. If mechanical energy leaks through the support-frame then the amplitude is reduced, but by supporting the system at a node losses are kept to a minimum (see Fig. 10). The Langévin design is a popular transmitter using nodal support, which also benefits from axial symmetry, allowing an axial screw to pull the vibrating components tightly into compression. Some materials used in electromechanical

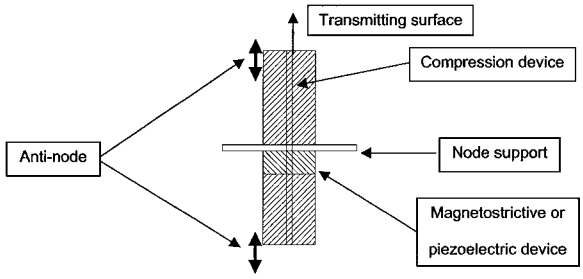


FIGURE 10 Sketch of a CW ultrasound source. The important feature is that the entire system vibrates so that the support is a node (minimum vibration level) and the two ends are anti-nodes (maximum vibration level).

devices are brittle, fracturing in tension but a static, axial compression force prevents tension developing in the device making it longer lasting and capable of working at higher amplitude. At resonance the two ends of the vibrating cylinder are anti-nodes, with a node at the center, making the cylinder a half a wavelength long. The frequency of resonance can be calculated knowing the speed of sound in the materials. Langévin resonators are used in sonar transmitters, inkjet printers, and sieve agitators.

B. Quality Assurance Using Resonance Spectroscopy

Resonance spectroscopy is a low power, CW application used for quality inspection. It is used to test the quality of mass-produced components because it is fast and the cost per test is low. A test sample is injected with CW ultrasound or sound at one transducer and the response at a second transducer is measured. A spectrum is built-up by stepping over many frequencies. Continuous wave excitation fills the sample under test with ultrasound, allowing waves from relatively distant parts of the sample to reach the receiver and contribute to the spectrum. A whole sample can be tested without any scanning and interpretation is done automatically by computer, typically using a trained artificial neural network, resulting in short testing times. The time to complete a full test of, 100 test frequencies on an automotive component could be as short as 1 sec so that more than 1000 parts per hour can be tested. Resonance spectroscopy is suitable for 100% quality assurance of mass-produced parts.

C. Measurements Using Phase

The phase of the received signal can be measured relative to the transmitted CW signal in a pitch-and-catch test, instead of using pulses. Small changes in the time of flight can be measured this way, which can form the basis of a useful quality test. The disadvantage is that phase measurements become ambiguous if the phase change is greater than one whole cycle (2π).

D. Noise Canceling

Destructive interference between two ultrasonic or acoustic waves reduces the amplitude of the resulting wave. Acoustic noise in a chosen region of space can be reduced or eliminated using destructive interference. An example is the cockpit of an aircraft, in the region of the pilot's head. Signals heading into the region of interest are sampled using microphones, inverted and retransmitted to interfere with the incoming waves. Noise reduction of between -6 and -24 dB can be achieved.

IV. MUSICAL INSTRUMENTS

Musical instruments all work in a band of frequencies occupying the range of frequencies from approximately 100 Hz–1 kHz. This is not the full range of human hearing but it is the range used for human speech. All traditional musical instruments and the human voice operate on the same principle—there are two components, a source of sound and a resonant cavity. The source of sound must have appreciable acoustic energy in the frequency range of the resonant cavity so that it can pump the resonator. There are two common ways of exciting the sound: impulse excitation (plucking, bowing, and striking) and causing air to move and oscillate (for example, a vibrating reed). There are only two forms of resonant cavity of importance: air in a cavity (wind instruments) and a solid vibrating object (stringed instruments and percussion).

A. Musical Scales

The term *pitch* is used in music to refer to the frequency of a musical note and it is internationally agreed that the musical note known as Treble A should have a frequency of 440 Hz for concerts. However, the absolute value of frequency is probably less important than the relative frequency between notes in terms of subjective musical appreciation. Relative frequencies in simple ratios lie at the heart of music, for example, an octave is double the frequency (2:1), the minor third is 6:5, the major third is 5:4, the fourth is 4:3, the fifth is 3:2, the minor sixth is 8:5, and the major sixth is 5:3. On a stringed instrument the octave is played by holding down a string at half its length.

B. Stringed Instruments

A stringed instrument comprises at least one string held under tension by a reaction frame. Strings are commonly made of polymers and metals and have some method of

adjusting tension, to adjust the pitch of the open-string. There is also a sound box, forming part of the reaction-frame, to provide a more efficient way of converting the energy of motion in the string into waves in air because a string without a sound box is virtually inaudible. Sound boxes also impart a tonal quality to the instrument, perhaps the best known examples are the violins made by Stradivarius. In recent years, however, electrical amplification methods have made sound boxes unnecessary and now some guitars and violins are made that can only be used with electrical amplification.

The rosin used on bows in the violin family of instruments gives a high coefficient of friction during sticking contact with the string and a low coefficient during sliding. The bowing action causes the string to be displaced into a triangular shape, with the apex of the triangle under the bow, then released. The equation of motion of the string is the wave equation in one dimension. A linear polarization solution simplifies the analysis and may be appropriate because the bow constrains the motion to lie in one plane. The speed of travel of transverse waves on the string is $c = \sqrt{(T/\rho)}$, where T is the force of tension in the string and ρ is the mass per unit length. Figure 11 shows the Fourier series solution plotted at various times in a cycle. The solution shows that most of the energy is concentrated in the low frequency modes of vibration.

The solution can be written as the sum of two traveling waves, with terms like $\sin(n\pi(ct - x)/L)$ and $\sin(n\pi(ct + x)/L)$, where n is an integer that identifies each mode, L is the length of the string, and x is a distance along the string. These waves can be seen in Fig. 11 most clearly at times $\pi/2$ and $3\pi/2$ as two triangular profiles, traveling in opposite directions, that are inverted and reflected when they reach an end of the string.

C. Wind Instruments

In this category of instrument are woodwind, brass, organ, and pipes. The source of sound can be a vibrating element, such as a reed in an oboe, or it may be the air vibrating, due to turbulence from a sharp edge as in a flute. Changing the resonating cavity, usually by altering the effective length, creates different notes. The clearest example is a church organ in which there are many pipes of different lengths. An organ pipe has an interesting set of boundary conditions: at one end the pipe is closed, creating a nodal point, at the other end it is open, creating an anti-node. The pattern of nodes forces the pipe, of length L, to be either a half a wavelength, $\lambda/2$, or an integer number and a half of wavelengths.

$$L = (2n + 1)\frac{\lambda}{2} \quad \text{where } n = 0,\ 1,\ 2,\ 3, \ldots$$

Given the speed of sound in air (330 ms^{-1}) it is possible to calculate that a pipe tuned to one octave below middle C (132 Hz) will be 1.25 m long. The next harmonic, the third, is at a frequency of 396 Hz. Organ pipes only create odd harmonics, which contributes to their characteristic sound. Woodwind instruments and flutes have open-ended resonators with several holes along the length. The holes can be opened or closed by the player, altering the effective length of the resonator and changing the note. The lowest note is always achieved when all the holes are closed and the resonator length is a maximum.

D. Percussive Instruments

The characteristic of percussion instrument is that they all use impulses to excite resonators, examples include the cymbals and the xylophone. The impulse provides energy and the resonator tunes the sound. The sound from cymbals covers a wide frequency range and is the closest to white noise that any orchestral instrument makes. An impulse contains energy in a wide frequency range but, unusually, the cymbal resonator is not tuned to one frequency, instead it can vibrate in a wide range of modes simultaneously giving its characteristic sound. Drum skins are also capable of multimodal vibration patterns. The perimeter of the drum skin is always a node but there can be a wide range of patterns of nodes and anti-nodes over the skin. At the fundamental note the center of the drum skin is the single anti-node and this note contains the majority of energy. The pitch depends upon the tension in the skin, its density, and the diameter of the drum.

E. Electronic Instruments and Composition

It has become popular to use signal processing for creating music. Either the acoustic signal from a conventional instrument or a purely synthesized signal can be used as the starting point for creating a new sound but, eventually, the signal will be amplified electronically and converted to sound by a loudspeaker. There is no need for a sounding box on an electronic musical instrument. A keyboard can be interfaced relatively easily to this technology and steel-stringed instruments can be adapted too because magnetic sensors (pick-ups) can detect the motion of the strings. A wide range of effects can be created, for example, harmonics can be filtered or added to simulate conventional instruments and echo can be added to create reverberation. It is possible to replace percussion instruments by using automatic rhythm and drum-beats. Multitrack tape recording and computers can control and play music that is too fast and complex for musicians, an extreme example being the playing of birdsong after

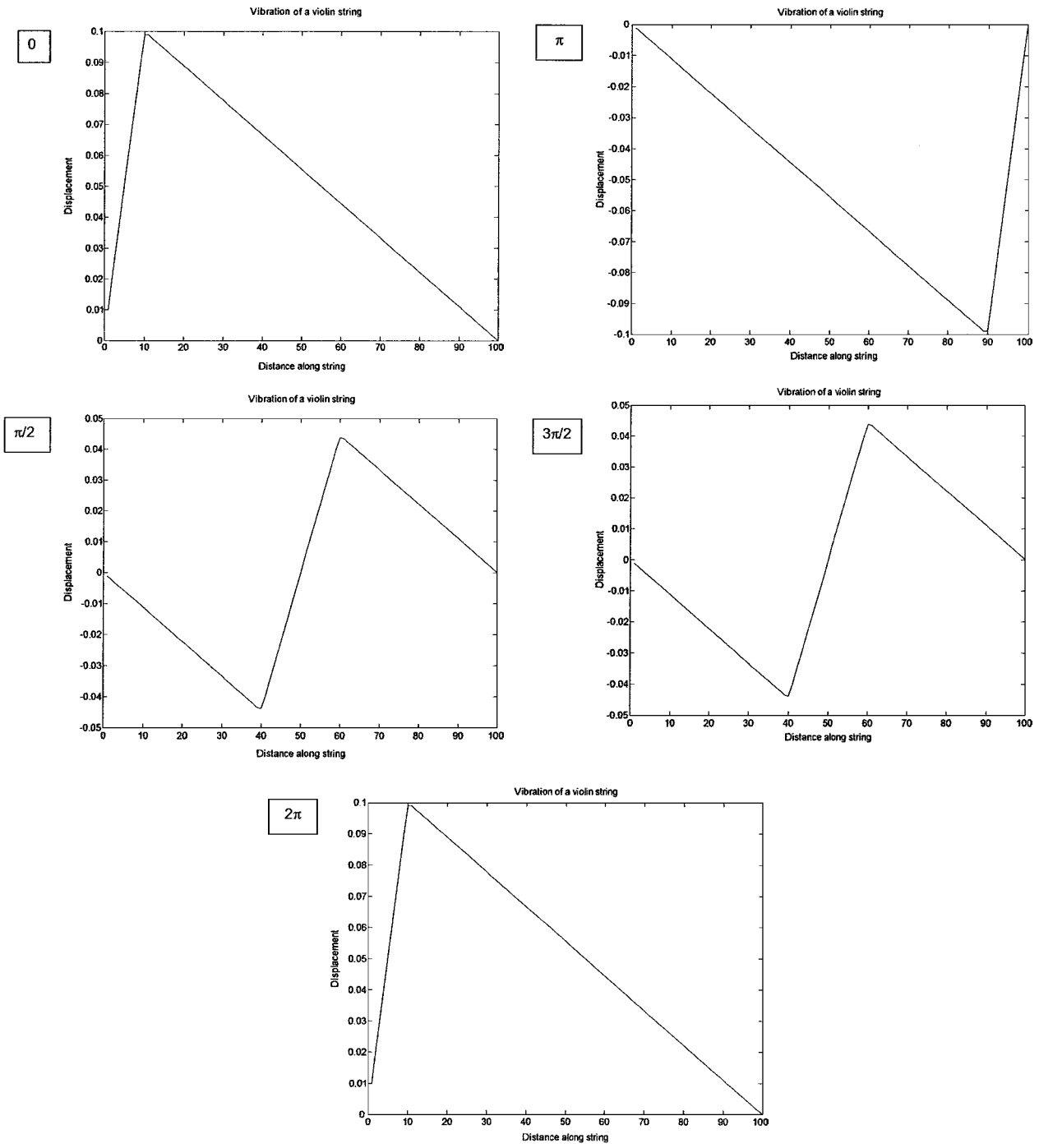


FIGURE 11 The Fourier series solution for the linearly polarized, transverse vibrations of a bowed string is evaluated at five times during a cycle. Displacement has been exaggerated. The string returns to its starting shape after one full cycle (2π) due to the absence of an energy-loss mechanism.

transcription to musical notation. Multitrack recording allows instruments and vocalists to be recorded individually, perhaps at different times and in different locations, before the final musical piece is mixed to achieve a pleasing balance.

V. TRANSDUCER ARRAYS

An array is a collection of two or more transducer elements working in concert. An array can be used by transmitting the same signal from the elements but with different time

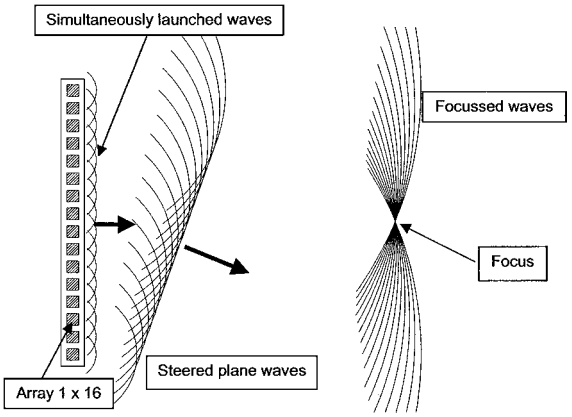


FIGURE 12 Sketch to illustrate how an ultrasonic beam can be steered and focused using an array of transmitters.

delays that change the direction of propagation of the transmitted wavefront or focus it. Arrays offer considerable flexibility in beam-forming, with no moving parts and high-speed scanning.

A. Steering and Focusing

Figure 12 shows groups of single wavefronts emerging from a linear array of elements. Where the wavefronts overlap with the same phase there is constructive interference but elsewhere the interference is destructive (not shown in the figure). It is common to space elements periodically, with a distance between centers of $\lambda/2$. This gives a good balance between source strength and beam quality. If elements are separated by more than $\lambda/2$ then there will be aliasing, called grating-lobes in this instance, in any wavefronts focused down as small as λ, meaning that the full potential for axial and lateral resolution available at the operating frequency is not achieved.

The same principles, of successive time delays and superposition of signals, can be applied to signals received by the array. A disadvantage is that a considerable amount of signal processing must be done. Beam-forming of received signals can be used to simulate the effect of an ultrasonic lens and beam-forming can provide better quality images than a lens, which is a further benefit of using arrays.

B. Time-Reversal Mirrors

Arrays can also be used in a self-adaptive imaging mode called time-reversal mirrors. First, a single impulse is transmitted simultaneously from all the elements in an array and the same elements are used for receiving. Then the received signals are simply time-reversed and retransmitted from the same elements. The elements are used again to collect a second set of signals, which is used for making a compound B-mode scan or C-scan (see Section II.D). Time-reversal mirroring automatically adapts the transmitted signals to the test sample, providing an image of any scatterers there.

C. Two-Dimensional Arrays

An array with elements distributed over two dimensions gives beam control in three dimensions. Arrays with 64×64 (4096) elements have been built but it is not feasible to transmit and receive from all the elements, instead only 128 receivers and 128 transmitters might typically be used in what is known as a sparse array. Since there are more than 256 elements available it is possible to use different elements as transmitter and receivers, but there is a very large number of ways to choose the elements, too many for all of them to be evaluated in a reasonable time. Computer simulations can select configurations that give good performance. Two-dimensional arrays allow arbitrary cross-sectional images to be created rapidly, giving better performance for medical imaging.

VI. MEDICAL ULTRASONICS

Applications of both pulsed and continuous (CW) ultrasound are to be found in medicine. Pulsed ultrasound is used for diagnostic imaging and CW is used for therapy purposes, for stimulating the healing of soft-tissues. One exception is a therapy instrument, using focused, pulsed ultrasound of high intensity to break stones in the body, for example, kidney stones. Frame-speeds of at least 10 frames per second are desirable for medical imaging to give the impression of a moving image. High-speed scanning is best achieved using electronic beam steering from arrays. One-dimensional arrays (1×128 transducers) or one and a half dimensional arrays (6×128 transducers) and two-dimensional arrays (64×64 transducers) are

TABLE VI Material Properties for Various Types of Human Tissue

Material	Wave speed (ms^{-1})	Acoustic impedance ($kg\,m^{-2}\,s^{-1}$)	Attenuation ($dB\,m^{-1}$)
Air	330	400	1200
Blood	1600	1.6×10^6	180
Bone	4100	7.8×10^6	2000
Brain	1600	1.6×10^6	850
Fat	1500	1.4×10^6	630
Kidney	1600	1.6×10^6	100
Liver	1600	1.7×10^6	940
Water	1500	1.5×10^6	22

used. The one-dimensional arrays are only capable of generating a single sector scan, a slice centered on the array. A two-dimensional array can scan a reasonable volume in real-time, for example, the heart or a fetus. It can provide data for an image plane of arbitrary orientation or for reconstructing a three-dimensional image.

Medical imaging equipment typically operates at center frequencies in the range 1–5 MHz. The transducers are damped and create short pulses of between $1^1/_2$ and $2^1/_2$ cycles, with axial resolution is in the range 5–0.5 mm. Higher frequency (10 MHz) transducers have been developed to detect abnormalities of the skin. Contrast in an ultrasound scan is due primarily to reflection of waves. Soft tissue and bone have different values of acoustic impedance therefore bones show up clearly in an ultrasound scan. A fetal skeleton, for example, has a high contrast allowing simple checks to be made on its growth and development during pregnancy.

The intensity of ultrasound used on fetuses must be carefully controlled. High intensity can cause tissue damage by cavitation (see Sections X and XI) and by heating. Nonlinear effects in water cause ultrasonic pulses to become sharper as they travel, increasing the intensity at the leading-edge of a pulse and another area of concern is that most imaging systems use arrays to focus the ultrasound—further increasing the intensity. Transmitter drive levels are kept low for safety. A transducer should be tested and certified before it is approved for use on fetuses. During testing its ultrasonic output is measured using a receiver which has a traceable calibration to a standard. An exceptionally small receiver (1-mm diameter) is needed, which is capable of working over an exceptionally wide range of frequencies (up to 100 MHz); one popular receiver satisfying these requirements is called a membrane hydrophone because it is made from a thin membrane of ferroelectric polymer (PvDF). Photographic methods are used to print fine electrode patterns onto the membrane and the size of these printed electrodes (0.1–1.0 mm) defines the lateral resolution of the hydrophone. The membrane is typically 0.1 mm thick, which defines the axial resolution.

VII. NONDESTRUCTIVE EVALUATION

Nondestructive evaluation covers a wide range of applications, such as: testing pressure vessels, testing welded joints, finding delamination flaws in aerospace assemblies, finding faults in silicon wafers, testing railway track for cracks, and testing gas turbine engine blades for casting faults. While the range of applications is great rather few different techniques are used. Applications can be classed as either flaw-detection or material property assessment. Transmission testing is commonly used for material property assessment and pulse-echo methods are used predominantly in flaw-detection with A-mode scans. C-scanning is used sometimes on samples that fit into water immersion tanks. Arrays are not used as frequently as in medical imaging. Heterogeneous materials such as austenitic stainless steel, cast iron, and concrete are considered particularly difficult to test because these materials can generate random multiplicative noise. It is, however, possible to detect line scatterers in concrete, such as reinforcement bars, at a range of 0.5 m. Other methods that are infrequently used are resonance spectroscopy (see Section III) and acoustic emission (see later). These are mentioned because the principles of operation are significantly different to the pulsed methods otherwise used in nondestructive evaluation.

A. Detecting and Sizing Flaws

Flaws reflect ultrasonic wave energy and generate echoes. Flaws can be detected, provided their echoes can be resolved from other echoes. Lateral resolution is provided by scanning a transducer over the surface of the test sample. The position of any flaw can be found by knowing, first, the position of the ultrasonic beam and, secondly, the time of arrival of its echo—from which the position of the flaw along the beam can be calculated knowing the speed of sound. The size can be found using waves diffracted from the tips of flaws, as illustrated in Fig. 13, which shows ultrasonic waves in glass passing, reflecting, and diffracting from the tip of a narrow slit viewed parallel to its plane. Compression waves are partially converted into shear waves, whose existence can be used to detect the extremities of the flaws.

B. Measuring Material Properties

Empirical relationships have been developed between material properties, such as attenuation and the speed of sound. Typical properties of interest include crushing strength (concrete), porosity (ceramics), and grain size (austenitic steel and cast iron). Experiments are nearly all performed in transmission, using pitch-and-catch, through a known distance in the test material. Many of these materials are heterogeneous, for which attenuation measurements can be unreliable (see Section I.G).

C. Acoustic Emission

Acoustic emission is a passive monitoring system used to give a warning of significant structural or material change. It is commonly used for monitoring machinery and structures during normal operations over long periods of time. There are no transmitters of ultrasound in an acoustic

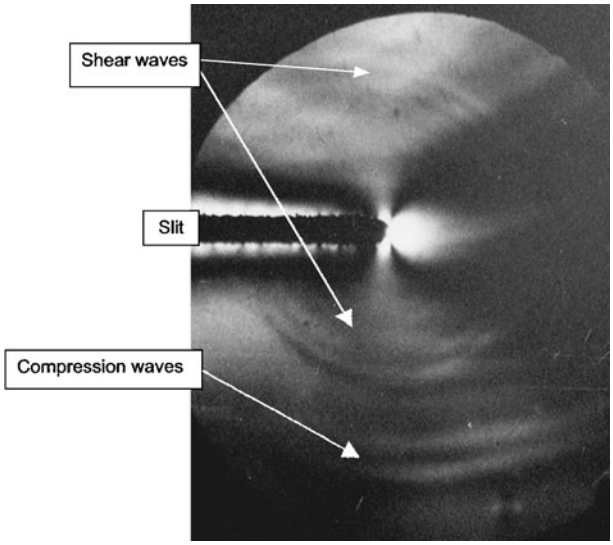


FIGURE 13 Planar compression waves (1 MHz) traveling from top to bottom in a glass block have been rendered visible passing a planar slit. The tip of the slit acts as a source of cylindrical shear waves, which travel more slowly than the compression waves.

emission system, only receivers. The source of sound is the machinery or structure under test. Ultrasound is created as a crack expands or as wearing progresses and the presence of ultrasound indicates that the material has been degraded. The location of the source can often be identified either from one receiver out of several or from cross correlation and triangulation. It is also sometimes possible to interpret the acoustic signal as the signature of a particular part wearing. Acoustic emission can be sensitive to environmental noise, for example, acoustic emission has been tried in a pilot scheme to monitor offshore oil production structures, but it proved difficult or impossible to separate material emissions from the noise made by sea waves washing against the structure. However, acoustic emission works well for testing components by exploiting the correlation between loading and cracking activity, for example, in testing turbine blades for gas-turbine engines, by loading the blades while simultaneously monitoring their acoustic emission it is possible to detect the presence of flaws—quiet blades pass the test but noisy blades fail.

VIII. GEOPHYSICAL EXPLORATION

The aim of geophysical exploration is to produce cross-sectional images of the earth showing the rock strata so that promising locations for drilling test wells can be identified. Pulsed sonic methods are used over the frequency range 0.1–100 Hz, with a single point source as a transmitter and a linear array of receivers. In a typical test a small explosive charge, or *shot*, is used as the source of sound with about 50 receivers or geophones. A narrow bore-hole is drilled about 100 m down into the underlying rock, into which is put the shot. The geophones are usually spaced 10 m apart or more and buried along a straight line (the *profile*). Recording equipment stores the signals from the geophones. A full analysis is done using computers to process the signals.

The basis of much of the interpretation assumes that the earth is a layered material, with each layer having its own material properties. The different layers reflect waves back to the surface, reverberate, and guide the waves. The main objective of processing the signals is to produce an image of the strata based upon the reflected data only. Other objectives are to compensate for amplitude reduction with range, to compensate for attenuation losses and to compensate for dispersion. Signals are displayed as either compound A-mode (or a variant of it known as VAR) or compound B-mode. The disadvantage of A-mode is that the image is difficult to understand and in B-mode phase information is lost. In VAR one half of the phase is blackened and this gives an image that is more easily understood. Many of the approaches used to process the signals are essentially the same as methods used elsewhere in the field of ultrasonics and acoustics. One example of this is stacking the signals from certain geophones. Stacking is used partly to average signals, thereby improving the signal-to-noise ratio, partly to perform spatial compounding, to help suppress the multiplicative noise created by random scatterers in the rock, and partly for focusing (common depth point, CDP, and common reflection point, CRP) and is therefore related to synthetic-aperture focusing used in sonar and radar. Deconvolution of the sonic pulse is also frequently applied (see Section II). Controlled chirp signals are sometimes used as an alternative to an explosive shot, when large motorized vehicles carrying heavy vibration units are used to transmit low-frequency, linear, sweep frequency chirps (0.1–10 Hz) into the ground. Another approach to interpreting seismic profiles is to predict a synthetic seismogram, using a model of the rock formations as a starting point. A computer program, much like a finite element program, then models the propagation of the sonic pulse through the rock strata. Where the predicted and experimental seismograms disagree it is possible to modify the model and re-predict the synthetic seismogram until the degree of disagreement is acceptably low.

IX. MARINE SONAR AND SONAR IN ANIMALS

Sonar systems can be classed as either active or passive. In active sonar an ultrasonic or acoustic pulse is transmitted

and echoes are detected. In active marine sonar echoes are displayed in A- or B-mode. In a passive marine sonar no signal is transmitted, instead the noise emitted by a target is monitored at several receivers, with which it is possible to measure differences in the times of arrival and calculate the bearing and range. It is common in the field of marine sonar to refer all pressure measurements to the level of 1 μPa with a corresponding reference intensity of 0.67×10^{-18} W m^{-2} in water.

$$\text{Intensity} = (\text{mean})\,\text{pressure}^2/\text{impedance}$$

A. Marine Sonar

Marine sonar can be used for warfare, sea bottom sounding, mapping the sea bottom and locating shoals of fish. Complications arise when working in deep oceans because the water temperature changes with depth, causing ultrasonic waves to be reflected. If a vessel moves slowly in a straight line, periodically collecting side-scan sonar signals it is possible to combine them into a compound B-mode image. The signals used in forming this image can then be processed using an algorithm called synthetic aperture focusing, which improves the quality of the image by synthesizing a large aperture lens to focus the image. Common working frequencies for side-scan sonar are in the range 100 kHz–1 MHz. Sonar used by warships makes use of passive and active array methods: transducer panels are attached to the hulls of ships, long linear arrays are towed behind ships, and helicopters lower sonar systems into the water while hovering. The return signal contains information about the target that can help identify it because all structures, including ships and submarines, respond to active sonar by resonating and re-radiating ultrasound (see Fig. 3B). Passive signals can be similarly analyzed. Passive countermeasures have the objective of making the submarine quiet and difficult to detect, for example, quiet propellers and engines, streamlined hulls, and surface treatments giving low ultrasonic reflection. Active countermeasures aim to confuse the location and identification of the target, for example, jamming sonar. Sonar systems used in warfare must detect targets at a reasonable distance and this constrains the working frequency range to about 1–100 kHz, with a typical wavelength of 1.5 m–1.5 cm. At the lower frequencies the greatest range is achieved, but it is then difficult to create a narrow beam because large transducer arrays are required (between 10 and 100 m). An interesting solution is to make use of nonlinear effects in the water. A parametric source is created in this way by driving a sonar transmitter at high power at two frequencies, f_1 and f_2. The nonlinear effects create two new waves, one of which is at the difference frequency ($f_1 - f_2$). Sources of difference frequency waves are created at many places in a long, periodic, linear array (a parametric array) in the water ahead of the ship. The beam is exceptionally narrow and points in the direction of the main sonar beam. The difference frequency is typically 10 kHz but it can be changed considerably by making relatively small changes to the two frequencies of the main sonar (approximately 100 kHz), so a parametric source is capable of considerable frequency agility, which is an advantage in view of the complexity of countermeasures used in military sonar systems.

B. Animal Sonar

Animals with particularly effective sonar systems include bats, whales, dolphins, and porpoises. These mammals have auditory systems similar to humans, with vocal chords to launch waves and two ears capable of phase discrimination so that the direction of a sound can be estimated. Some bats have sonar systems that work at frequencies up to 100 kHz. All bats use controlled chirp signals when hunting. A relatively short chirp with a wide bandwidth is employed when looking for prey, giving good spatial resolution. Once the bat has detected an echo it increases the repetition frequency of transmission and changes to a longer chirp with a narrower bandwidth, which is good for detecting a Doppler signal. Perhaps the bat can determine the relative speed and direction of its prey or perhaps it can identify the prey using this kind of chirp. As the bat approaches to catch its prey short, large bandwidth chirps are used again at the highest repetition rate for precise spatial information.

Mammals using sonar in the sea get information about the skeletal mass of the peers in their social groups. This is because the reflection coefficient from water to soft tissue is about 0.1% whereas the value for water to bone is almost 50%, so the skeleton generates the strongest echo signals. It is not known if the information is registered as an image in their brains; a close analogy, however, would be fetal imaging (see Section VI). Dolphins and porpoises use their sonar to find fish and they also use it to stun the fish, by emitting an intense burst of waves. Whales and dolphins use frequencies from about 1 Hz up to about 30 kHz for their sonar.

X. PROCESSING TECHNOLOGY

Ultrasound is used to cause agitation in liquids and powders in most of the applications of processing technology. Significant power levels are needed so generally continuous ultrasound is used. Ultrasound can be used to nebulize liquids, it can also be used to trap and separate particles in a standing wave where particles will move to the nodal

points in the pattern. High-power ultrasound, at a frequency of approximately 40 kHz, can be focused to a point using an acoustic lens, where the intensity can be sufficiently high to weld plastics. Similar focusing devices have been used for de-fouling of organic material in underwater applications. Medium-power ultrasound is used to agitate sieves, increasing the rate of powder sieving in the manufacture of pharmaceutical products. Probably the most commonly used ultrasonic instrument in this field is the ultrasonic cleaning bath. It is a simple container, usually made of stainless steel, filled with a liquid, usually water, and agitated by one or more ultrasonic transducers attached to the outside of the bath. The working frequency range is 40–150 kHz. The drive frequency and amplitude can usually be adjusted, along with the cleaning time. Cleaning can be caused by motion of the liquid or by more aggressive cavitating bubbles, because when a bubble collapses it can generate very high temperatures and pressures and it can also create erosive, microscopic water-jets.

XI. SONOCHEMISTRY

The presence of an ultrasonic field can have the following effects upon chemical reactions: it can accelerate the rate of reactions that occur in its absence and it can enable reactions to occur that otherwise would not happen. The latter is classed as sonochemistry.

loosenes-1Sonochemistry is closely related to sonoluminescence, which is the emission of light when a liquid has an intense sound field in it. Sonochemistry occurs if there is an ultrasonic field with a frequency of approximately 300 kHz and the intensity is sufficiently great to cause cavitation. The liquid must also contain molecules of an inert, monatomic gas such as argon. When a liquid cavitates it forms vapor bubbles on the rarefaction half of the pressure cycle. The bubbles collapse quickly as the pressure changes to compression and this is believed to cause both sonoluminescence and sonochemistry. Bubbles collapse in about 1 μs and pressures and temperatures as great as 5×10^8 Pa and 10^4 K have been measured during collapse. Temperatures of 10^3 K would be sufficient to account for some of the sonochemical effects reported so sonochemistry and sonoluminescence are believed to be thermochemical effects, caused indirectly by the high temperatures in the collapsing bubbles. Sonochemical effects are generally absent without the presence of the inert, monatomic gas mentioned earlier. Monatomic gases have higher ratios of specific heats (1.33–1.67) than either diatomic or polyatomic gases, therefore, a collapsing bubble of a monatomic gas generates the highest temperature of any gas. Short duration spectroscopic analysis has shown there are generally two components in sonoluminescence: an infrared component associated with black-body radiation, allowing temperature to be estimated and spectral lines associated with excited states of some active chemicals. Pressure broadening of the spectral lines is the basis of the pressure measurement.

Sonochemistry offers some unique advantages to chemists: an average, operating temperature close to ambient, a high reaction temperature up to 10^4 K, precise control over the location of the reaction (the ultrasonic field), and a large number of disbursed, reaction sites (the cavitating bubbles), which should help to achieve high reaction yields.

SEE ALSO THE FOLLOWING ARTICLES

ACOUSTICS, LINEAR • ACOUSTIC WAVE DEVICES • EXPLORATION GEOPHYSICS • MUSICAL ACOUSTICS • SONOLUMINESCENCE AND SONOCHEMISTRY • WAVE PHENOMENA

BIBLIOGRAPHY

Auld, B. A. (1990). "Acoustic fields and waves," 2nd ed., Krieger Publishing Company, FL.

Birks, A. S., and Green, R. E., eds. (1991). "Nondestructive Testing Handbook," Vol. 7 Ultrasonic Testing, 2nd ed, Am. Soc. Nondestructive Testing.

Bushong, S. C., and Archer, B. J. (1991). "Diagnostic Ultrasound," Mosby-Year Book, St. Louis, MD.

Coates, R. F. W. (1990). "Underwater Acoustic Systems," MacMillan Education, Basingstoke,

Crampin, S. (1987). Geological and Industrial implications of extensive-dilatancy anisotropy. *Nature* **328**.

Fink, M. (1992). "Time reversal of ultrasonic fields—Part 1 Basic principles and 2 Experimental results," IEE Trans Ultrason., Ferroelect., Freq., Contr 39, No. 5.

Kino, G. S. (1987). "Acoustic waves," Prentice-Hall, New Jersey.

Leighton, T. G. (1994). "The Acoustic Bubble," Academic Press, San Diego.

Lynn, P. A. (1987). "Radar Systems," MacMillan Education, Basingstoke,

Maynard, J. D. (1996). Resonant ultrasound spectroscopy. *Phys. Today* **40**.

Vacuum Arcs

James M. Lafferty
General Electric Company, retired

GLOSSARY

Anode spot Molten area on the positive electrode surface of an arc that releases metal vapor because of heat generated by concentrated electron bombardment.
Arc Self-sustained, low-voltage, high-current electrical discharge.
Cathode spot Highly mobile, minute luminous area on the negative electrode surface of an arc that emits electrons, jets of plasma, and metallic vapor.
Cathode-spot track Erosion marks left on the negative electrode of an arc by the passage of a cathode spot.
Chopping current Magnitude of the arc current just prior to arc extinction.
Current chopping Sudden cessation of arc current at the time of arc extinction.
Current zero Zero of an alternating sinusoidal current.
Electron temperature Average translational kinetic energy of the electrons in a plasma.
Field emission Liberation of electrons from unheated metal surfaces produced by sufficiently strong electric fields.
Plasma Region in an electrical discharge that contains very nearly equal numbers of positive ions and electrons and may contain neutral particles as well.
Sheath Space-charge region at the boundary of a plasma with an excess of either positive or negative charges.
Thermionic emission Evaporation of electrons from a metal surface produced by heating the metal.

AN ARC may be defined as a discharge of electricity, between eletrodes in a gas or vapor, that has a voltage drop at the cathode of the order of the minimum ionizing or minimum exciting potential of the gas or vapor. The arc is a self-sustained discharge capable of supporting large currents by providing its own mechanism for electron emission from cathode spots on the negative electrod. However, the term *vacuum arc* is a misnomer. What is really meant is a "metal vapor arc in a vacuum environment." But since vacuum arc is in common usage and has been accepted in the literature, it is retained here. A vacuum arc, then, burns in an enclosed volume that, at ignition, is a high vacuum. A characteristic feature of such an arc is that after ignition it produces its own vapor from the electrodes that is needed to achieve the current transport between the electrodes. In such a vacuum arc, one can clearly distinguish phenomena that occur at the cathode and the anode and in the plasma occupying the space between the electrodes. Because these phenomena are exceedingly complex and interrelated, there are no general theories that completely describe the vacuum arc or predict its behavior.

The electrical discharge in a mercury-arc rectifier tube is an important example of a vacuum arc. Here the metal

vapor is mercury supplies mainly by the cathode spots on a mercury pool. An anode, usually made of carbon, collects electrons from the plasma. Evidence of vacuum arcs has been found on the walls of various fusion devices. Here one observes cathode spots and cathode-spot tracks. Such arcs are believed to be of the homopolar type. The discharge in a thyratron is also characterized by a low voltage and high current, but is not a true arc because of the absence of cathode spots. The energy to heat the thermionic emitting cathode is supplied by an external source, not by the discharge itself. The discharges found in vacuum circuit breakers and triggered vacuum gaps are true vacuum arcs. These are described at the end of this article.

I. GENERAL DESCRIPTION OF THE VACUUM ARC

The vacuum arc is shown in idealized form in Fig. 1. A slow-motion color movie of such an arc is a beautiful sight to behold. One sees a cold cathode surface covered with isolated, small, brilliant spots. These cathode spots move erratically over the cathode surface, sometimes dividing into two or more fragments or extinquishing and reforming elsewhere on the cathode. Associated with the cathode spots are luminous jets that shoot off into space and constantly change direction. Between the electrodes one sees a diffuse glow whose color is characteristic of the electronically excited metal vapor of the electrodes. If the arc has been burning for some time, one is also likely to see on the anode a bright stationary spot of molten metal. This anode spot will be completely surrounded by an intense glow. As one looks away from the central region of the arc, the glow becomes more diffuse, eventually disappearing. Let us now examine these phenomena in more detail.

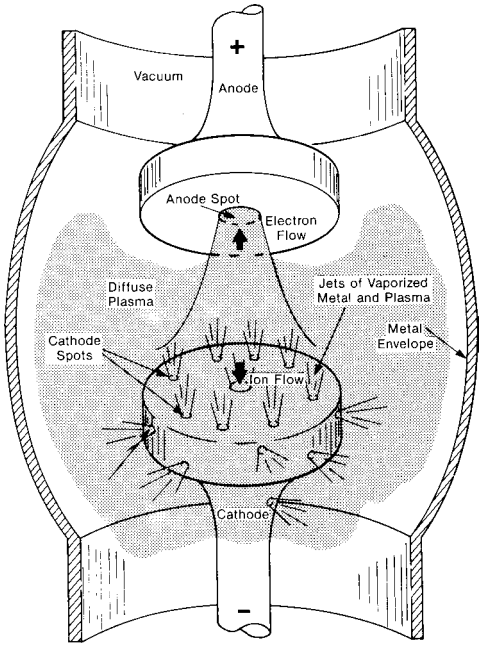


FIGURE 1 Structure of the vacuum arc.

The high-current vacuum arc forms in the metal vapor evaporated from the electrodes. Electron current is fed into the arc by a multiplicity of highly mobile cathode spots that move about on the negative electrode. The current density in these small spots is exceedingly high and is often of the order of a million amperes per square centimeter or more. Jets of plasma and metallic vapor, with velocities of up to 1000 m/sec, also have their origin at the cathode spots. In the formation of these jets, which are the principal source of plasma and vapor in the vacuum arc, one atom of metal may be removed from the cathode for every 10 electrons emitted. The moving cathode spots leave pitted tracks on the cathode surface that show little evidence of cathode melting, although these cathode-spot tracks often suggest loss of metal by sublimation. The exact mechanism responsible for the emission of electrons and the ejection of metal vapor and plasma from the cathode spots is not fully understood and has been a source of wonder and controversy since the phenomena were first investigated. However, thermionic emission and field emission undoubtedly play an important role, since positive ion bombardment and space-charge effects produce extreme local heating and intense electric fields, respectively, at the cathode surface.

Upon arc ignition (Section II.A), the space between the electrodes quickly fills with a diffuse plasma consisting of partially ionized metal vapor. At high currents this plasma expands into the volume surrounding the electrodes and their supports. At low currents the positive electrode collects electron current from the plasma uniformly over its surface. The metal shield or vacuum envelope that surrounds the arc also collects charges from the plasma and metal vapor. At high currents one or more distinct anode spots may appear. These spots always form on the end of the anode, which faces the cathode, in contrast with the cathode spots, which some-times wander off the end of the cathode and move about on the sides of the electrode. The anode spot also tends to remain in one position, in contrast with the mobility of its counterparts on the cathode.

At the anode, electrons striking the electrode carry essentially all of the arc current, whereas at the cathode, the ions striking the surface account for only about 10% of the current, with emitted elections supplying the remainder. Consequently, there is a difference in the power dissipated at the two electrode surfaces. The cathode is bombarded by ions with a relatively low total energy, but

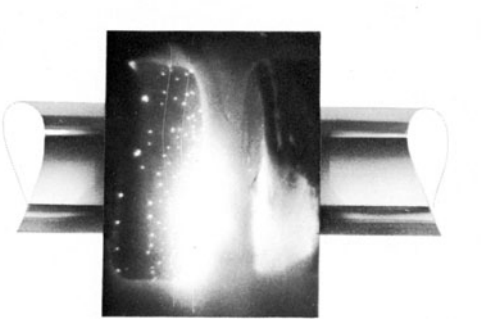


FIGURE 2 Photograph of a vacuum arc showing anode spot and cathode spots with plasma jets. (Photograph courtesy of Dr. Gerhard Frind, Corporate Research and Development, General Electric Company.)

since most of these are landing in the small areas of the cathode spots, the power density is high. At the anode, on the other hand, a large stream of high-energy electrons is bombarding the relatively large anode spot, subjecting the anode to a much higher total power input. This bombardment inevitably causes the anode surface to melt and vaporize at high currents. The magnetic pinch effect also contributes to a concentration of energy input to the anode spot. This concentration of power will eventually destroy the anode unless the arc is extinguished. Figure 2 shows a photograph of the typical vacuum arc with an anode spot.

II. DETAILED CHARACTERISTICS OF THE VACUUM ARC

A. Arc Ignition

The essential requirement of establishing an arc between metal electrodes in a vacuum is to initiate an electrical discharge that will lead to the development of a cathode spot. This development requires the presence of a plasma, that is, both electrons and ions, in the gap. The process usually begins by inducing cold-cathode electron emission, which builds up to the point where the heat generated vaporizes a minuscule portion of either or both electrodes. The metal vapor thus produced is partially ionized by the electrons. Positive ions and radiation then strike the cathode and enhance the electron emission. With the release of more vapor, these effects become cumulative. The discharge then develops into a metal-vapor arc, with intense ion bombardment of the cathode leading to the formation of a cathode spot. Additional cathode spots may then develop, depending on the magnitude of the arc current, which is determined almost entirely by the external circuit. It is essential that the power supply connected to the electrodes be capable of supplying a sufficiently large current to maintain a stable arc. The value of this current will depend on the electrode material but will range from a few tens of amperes to several hundreds, as discussed in Section II.E. The open-circuit voltage required to initiate the arc will depend on the triggering method. A simple method of starting the arc by increasing the applied voltage will now be discussed in some detail.

As the voltage across a vacuum gap is increased, electrical breakdown will eventually occur. At a certain level of voltage, in the range of many kilovolts, the negative electrode will begin to emit electrons. This room-temperature emission occurs as tiny jets form small regions on the surface, and the emission's magnitude increases sharply with further increases in voltage, leading ultimately to breakdown. This field emission is dependent on the work function and magnitude of the electric field at the cathode surface. The latter may be enhanced by surface roughness or protrusions on the cathode. The total emission current will generally be much less than an ampere; but since the areas of emission are small, the current densities reach extremely high values. As current densities in the 10^8 A/cm^2 range are reached, for a particular area, breakdown will generally occur. At these current densities joule heating occurs and the small emitter area is vaporized. The metallic vapor produced is partially ionized by the emitted electrons. The positive ions are accelerated by the electric field between the electrodes toward the cathode surface, which is bombarded with considerable energy. When this bombardment eventually causes the formation of a cathode spot, the vacuum arc is established. Additional cathode spots may then develop, depending on the magnitude of the arc current, which is determined by the external circuit.

Obviously, absorbed gas, dust, and especially insulating particles on the cathode will effect the breakdown voltage. In long gaps the positive electrode may also affect the breakdown process. In this case, high-energy electrons bombard the anode and produce metal vapor, which is then ionized. The ions thus produced strike the cathode and produce a more abundant electron emission. Under some circumstance, loose metallic particles may become charged and then accelerated across the gap by the electric field. On impact these particles may vaporize or produce vaporization of the electrode material. The presence of loose particles can reduce the breakdown voltage of a vacuum gap to a level as much as 50% below that of the particle-free case. There is also evidence that for a given electrode geometry and gap spacing, the breakdown voltage increases with the hardness and mechanical strength of the electrode material.

A vacuum arc may also be initiated with less than 100 V across the gap by first bringing the electrodes in contact and then separating them with current flowing. As the electrodes are parted, the area in contact diminishes

until only a small metal constriction bridges the two electrodes. This bridge is heated to the melting point by resistive heating, and vaporization follows at an explosive rate. The high temperature and electric field at the negative electrode cause a field-enhanced thermionic electron emission to flow. This emission ionizes the metal vapor, thus producing a plasma capable of carrying an arc current within the space formerly occupied by the bridge. Intense ion bombardment of the cathode leads to the formation of a cathode spot and the arc is fully established as the electrodes are further separated. A similar type of breakdown can be obtained by placing a wire between the separated electrodes and melting it with a high-current discharge.

In the presence of a high-electric-field coldcathode, electron emission may be initiated by injecting a plasma into a vacuum gap from an external source or by bombarding either electrode with high-energy electrons, ions, or radiation from, for example, a laser beam. Any of these processes will lead to breakdown and the formation of a vacuum arc.

B. Cathode Spots

The cathode spot, which is essential for the very existence of a vacuum arc, is the least understood of all vacuum-arc phenomena. Clearly, it is the source of electron emission for the arc, but it also provides plasma and metal vapor. The cathode spot is a highly efficient electron emitter. The current carried by a single spot depends on the cathode material and may vary from several amperes to a few hundred for most metals. When the arc current exceeds the current that a single spot normally carries, additional spots will form, sometimes by the enlargement or division of the original spot and also by the formation of new spots. Cathode spots do not respond instantly to a demand for an increase in current when the voltage applied across the arc is increased. It would appear that the spots normally operate at maximum current for the available heated emitting area and require a thermal response time of several microseconds to meet a demand for increased current.

Sometimes what appears to the eye as a single cathode spot is actually, on closer examination, numerous, small, active areas. This cellular substructure is found more frequently on mercury cathodes and may consist of a cluster of 4–12 cells.

Because cathode spots have a finite lifetime, they often extinguish and reform elsewhere on the cathode while the arc is burning. These spots are seldom stationary and move about on the cathode surface in a random, erratic way, sometimes reaching speeds of 30 m/sec on copper. It would appear that they tend to repel one another and also move in reverse to the direction expected for a conductor carrying current when a magnetic field is applied parallel to the cathode surface.

When the arc current is reduced, by decreasing the applied voltage or inserting resistance in the external circuit, the number of cathode spots will diminish. As the current is further reduced, a point is reached where only one cathode spot remains. The remaining spot has a statistical lifetime that is quite short at low currents and it will suddenly vanish during a period of less than a microsecond. When this occurs the arc is extinguished. The arc current flowing just prior to arc extinction is called the chopping current.

The current density in a cathode spot is phenomenally high. Estimates have ranged from 10^3 to 10^8 A/cm^2. This wide range of uncertainty is associated with the problem of accurately determining the active emitting area of a cathode spot that may range in size from 10 μm to several hundred. Regardles of the uncertainty, the electron emission current density is enormous compared to conventional electron emitters. The mechanism that provides such high emission densities is not well understood but is believed to be some form of field-enhanced thermionic emission.

In addition to the electron emission, the cathode spots also provide a copious supply of ions for the vacuum arc. Experimental analysis has shown that this stream of ions consists of singly and multiply charged high-energy positive ions of cathode metal. A majority of the ions have energies greater than the arc voltage and, consequently, would have no difficulty in reaching the anode. The cathode spot ions are rich in multiply ionized particles that frequently exceed the singly ionized ones, especially for refractory metal cathodes. The total ion current has been estimated at about 8% of the arc current.

Cathode spots are also a source of evaporated neutral metal vapor, and they eject tiny liquid droplets along trajectories nearly parallel to the cathode surface. Unlike the ions, the energies of most of the neutral atoms ejected are quite low, corresponding to the temperatures of the portions of the cathode surface from which they are evaporated. It is believed that most of the neutral vapor found in the vacuum arc does not come directly from the cathode spots themselves, but is evaporated from inactive former sites of the moving spots and from the micrometer-size liquid droplets ejected by the cathode spots during their flight to the electrodes and walls surrounding the arc.

The moving cathode spots leave behind on the cathode surface a trail of irregular pits, craters, and depressed valleys called cathode-spot tracks. Although these tracks usually show little evidence of gross melting, microscopic examination shows melting, evaporation, and possibly loss of metal by sublimation. The tracks are not always continuous and sometimes show areas that appear almost

undisturbed. Some areas of the tracks show evidence of splashing of liquid metal, leaving behind a trail of solidified waves and spherical droplets. The amount of erosion from these cathode tracks is usually expressed as micrograms of mass eroded per coulomb of charge passing through the arc. It is found to vary significantly with experimental conditions such as arcing time, arc current, and cathode size. For copper cathodes the evaporation loss is between 35 to 40 μg/C. However, there is a much larger loss of metal, nearly seven times, associated with the ejection of liquid droplets.

C. Anode Spots

The main function of the anode in a vacuum arc is to collect sufficient electrons from the ambient plasma to sustain the external circuit current. However, it is probable that some high-energy positive ions from the cathode spots also reach the anode. In addition, the anode receives neutral atoms of metal vapor that may condense on its surface and radiant energy from both the cathode and the plasma. All these quantities tend to heat the anode surface, but electron bombardment is by far the largest contributor. As a result, the anode will melt if the arc is allowed to persist for any appreciable time and may even occur within a few milliseconds at current levels of several thousand amperes. Continued heating of the anode leads to an arc instability that can be described as follows.

At low currents when the vacuum arc is first initiated, the anode collects electron current uniformly over its surface from the plasma. Normally, the anode is not a positive ion source and the small electric field in the plasma tends to drive the positive ions away from the anode. The absence of positive ions causes the buildup of an electron space-charge sheath around the anode with an accompanying anode drop of potential through which the electrons are accelerated as they stream to the anode. These electrons bombard the anode with energies that not only correspond to the anode fall of potential, but to the average electron temperature energy and the electron heat of condensation as well. At a sufficiently high current density, some local area of the anode—usually an edge (Fig. 2) or a rough spot where the heat conductance is poor—will be heated enough to release metal vapor. The vapor is immediately ionized by the high-energy electron stream. The positive ions neutralize the negative space charge in the volume adjacent to the anode area emitting the metal vapor and produce a sharp reduction in the anode fall of potential. The lower voltage drop causes more current to flow into the anode in this area. This increase in current flow is also aided by an increase in the random electron current density in the plasma resulting from the increased metal vapor density. The local increase in current flow to the anode heats it even more, causing additional metal vapor to be emitted and ionized. This condition leads to a runaway effect, causing a constriction of the arc at the anode with the production of an anode spot. At this point, the total arc voltage will decrease. The intense local heating produced at the anode spot may cause destructive melting of the anode unless means are provided to disperse, or force rapid motion of, the anode spot. Unlike cathode spots, anode spots readily respond to the force of a magnetic field applied parallel to the anode surface and move in the direction expected.

The temperature of the anode spot for copper electrodes has been estimated to be between 2500 and 3000 K. Temperatures are lower for the more volatile metals and higher for the more refractory ones. Unlike cathode spots, the anode spots are quite diffuse with areas up to a square centimeter depending on the arc current. Current densities of 10,000 A/cm^2 are typical in an anode spot. Since the formation of an anode spot is heat induced, it is clear that points, irregularities, sharp edges, and loosely attached particles can cause variations in the spot initiation. It is not uncommon for more than one spot to form simultaneously on an anode, but these usually draw together and collapse into a single anode spot. Luminous areas have been seen on the anode in vacuum arcs that are not true anode spots, especially in low-current arcs and arcs of short duration. Instead, these may be anode spots in the early stages of development or other inhomogeneities in the arc or on the anode surface. When a true anode spot fully develops, an anode jet is clearly visible and there is a copious supply of vapor to the discharge. The arc then constricts and there is a substantial drop in arc voltage. The arc becomes more stable and takes on the characteristics of a high-pressure arc, which indeed it may well be since the metal vapor density is approaching that of atmospheric pressure.

D. Interelectrode Phenomena

The processes that occur in the space between the cathode and anode of a vacuum arc are rather unevenful compared to those at the electrodes, but nonetheless they are very complicated and not easily subject to analysis. The interelectrode volume is filled with a diffuse plasma whose main function is to provide a conductive medium for the arc current transport between the electrodes. In the absence of an anode spot, the vapor and most of this plasma is provided by the spray from the high-velocity jets that have their origin in the cathode spots.

The voltage distribution in a vacuum arc has been idealized in Fig. 3. At the cathode, one must distinguish between the potential distribution in the region of the cathode spot and the areas away from the spot. The solid line

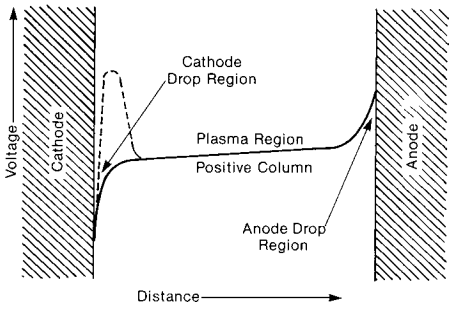


FIGURE 3 Schematic representation of the potential distribution in a vacuum arc.

showing the potential varying montonically between the cathode and anode is the classical distribution that one would expect in an area remote from a cathode spot. This solid line shows a region close to the cathode with a potential drop roughly equal to the ionization potential of the electrode metal, normally about 8V. The dotted line shows the potential distribution that may exist in a cathode spot region. This voltage hump has not been verified experimentally but is predicted on the basis of experimental observations that singly and multiply charged ions with energies in excess of the arc voltage exist in the interelectrode plasma. The details of this voltage distribution are unknown, and it may not even exist since the presence of the high-energy ions has also been explained to be the result of magnetic compression and heat expansion processes in the cathode spot.

Beyond the cathode fall region lies the positive column. In this region there exists a highly ionized diffuse plasma at a temperature approximating 10,000 K. The potential gradient is quite small because the ions and electrons are present in essentially equal numbers and because the conductivity of the plasma is very high, approaching that of a metal. The potential drop across the positive column will depend on the arc current and the column length, which is usually quite short for vacuum arcs and is only a few volts for arc currents of several hundred amperes or less.

The plasma and metal vapor in the positive column extend out to the shield or container walls encompassing the arc. The metal vapor and metal droplets from the jets will collect on the relatively cool walls. In addition, if the shield is not electrically connected to the arc circuit, it will assume a floating potential that is slightly negative with respect to the plasma and collect ions and electrons at the same rate. If the shield is connected to the cathode, a positive ion sheath is formed over its inner surface adjacent to the plasma. This is in contrast to the flaming sheath that surrounds a high-pressure arc operating in open air.

Before the formation of an anode spot, the metal vapor and plasma density in the positive column are so low that the mean free path of the electrons is comparable to the dimensions of the arc. Thus, thermal ionization and space-charge neutralization must be minimal. However, in the region of the cathode spots the metal vapor density is very high and ionization can readily occur.

The voltage drop at the anode is a result of negative space charge produced by the flow of electrons from the plasma to the anode in the absence of positive ions. Normally, the anode is not a source of positive ions and the potential gradient in the positive column tends to drive the positive ions away from the anode. The magnitude of the anode fall of potential will depend on the arc current. Once the anode is heated to its vaporization point, positive ions will form in this space and the potential drops drastically.

E. Arc Stability

All arcs tend to be unstable, particularly at low currents, because the cathode spots on which the arc is dependent for its existence are inherently unstable. Without knowing the intricate details of all the phenomena that occur in cathode spots, one can make some general observations about their behavior. Obviously, the cathode spot requires a feedback mechanism for its existence. The cathode-spot surface emits electrons and neutral vapor. These must interact to produce positive ions, which, in turn, bombard the cathode surface to produce the required temperature and field for additional emission of electrons and vapor. If this feedback becomes less than unity, the cathode spot will cease to exist. By observation of the fanatical motion of cathode spots over all faces of the cathode surface, with complete disregard to the position of the main discharge, one can only conclude that the cathode spot feedback mechanism is only slightly greater than unity and the spot is continually seeking a favorable environment on the cathode surface for its survival. Further evidence of cathode-spot instability is given by their short average statistical lifetime. For a copper cathode spot, the lifetime may average on the order of only a few tens of microseconds when the arc current is limited by the external circuit to a few amperes. The life increases markedly with current, and for a 10-fold increase in current, the lifetime increases 10,000 times.

If an adequate power supply voltage, well above the arc voltage, is assumed, high-current vacuum arcs of 1000 A or more will burn indefinitely. The burning time is limited only by the ability of the electrodes to dissipate the heat. Even though the lives of the individual cathode spots may be quite short, there are always enough of them in existence at any one time in a high current to keep it burning.

At arc currents in the 100 A range when only a few cathode spots may be present, the probability of them all vanishing at the same instant is quite high, and when this occurs the arc is extinguished.

Vacuum arc lifetimes not only increase strongly with increasing arc current, but are also dependent on the properties of the electrode material. Average lifetimes tend to increase with increasing atomic weight and vapor pressure and to decrease with increasing thermal conductivity. Figure 4 shows the average lifetime of arcs drawn between electrodes of various metals in vacuum as a function of arc current.

Increasing the circuit voltage above the arc voltage has only a modest effect on increasing the arc lifetime. The same is true for increasing the series inductance of the circuit. On the other hand, increasing the parallel capacitance across the arc will shorten the life of a vacuum arc. There is evidence that low-current arcs may survive by a succession of extinction–re-ignition events during which transient re-ignition voltages are developed by the series inductance of the circuit and the rapid decrease in current by the abrupt extinction of the arc. Excessive capacitance in parallel with the arc will limit the rate of rise of this voltage, thereby reducing the probability of arc re-ignition.

Unlike many gas discharge devices, vacuum arcs have a positive resistance characteristic; that is, the arc voltage

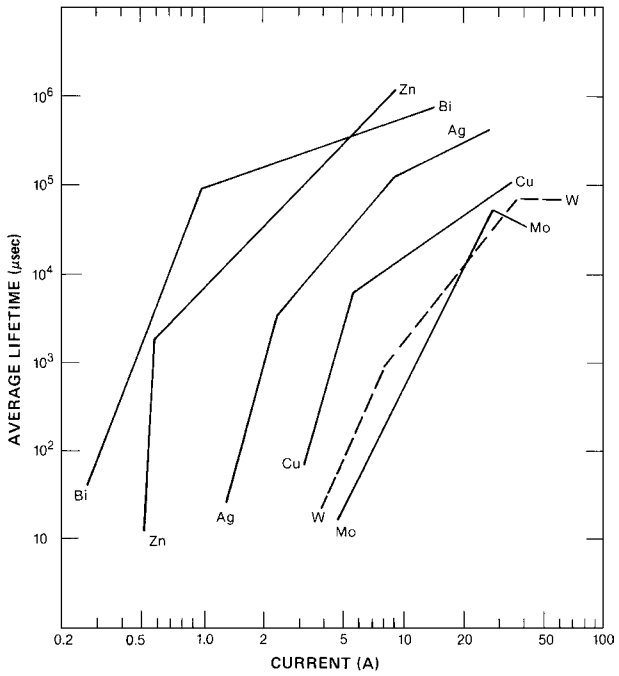


FIGURE 4 The average lifetime of low-current vacuum arcs for various electrode metals. [Reproduced with permission from Farrall, G. A., Lafferty, J. M., and Cobine, J. D. (1963). *IEEE Trans Commun. Electron*. **CE-66,** 253. © 1963 IEEE.]

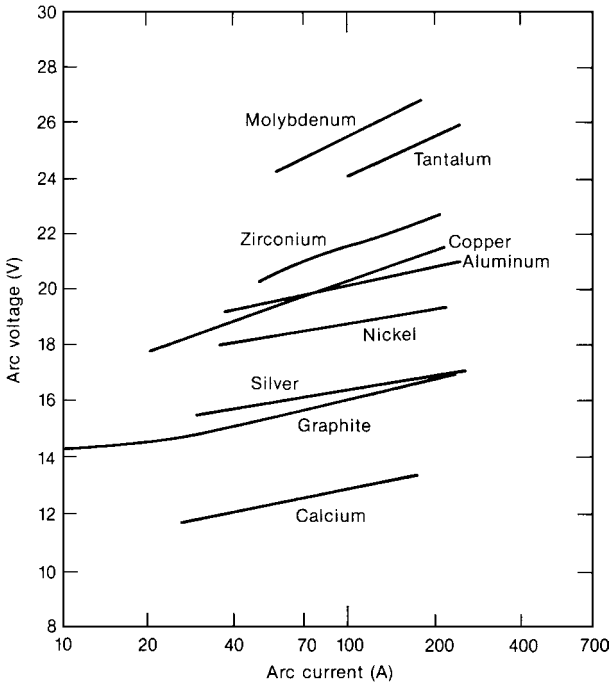


FIGURE 5 The volt–ampere characteristics of vacuum arcs for various electrode metals. [Reproduced with permission from Davis, W. D., and Miller, H. C. (1969). *J. Appl. Phys*. **40,** 2212.]

increases with arc current. Because of this characteristic, two or more vacuum arcs will operate stably in parallel without the need for series ballast impedances. Figure 5 shows the volt–ampere characteristics of low-current vacuum arcs for a number of electrode materials. The electrodes are approximately 1.27 cm in diameter separated by 0.5 cm.

Even at low-arc currents where only one cathode spot exists, the arc still exhibits a positive resistance characteristic. One can thus assume that a single cathode spot also has a positive resistance characteristic, as indeed must be the case since multiple cathode spots exist in parallel in high-current vacuum arcs. Probe measurements indicate that most of the increase in arc voltage with increasing current occurs at the electron space-charge sheath at the anode.

Even though vacuum arcs are stable at high currents, they, like most gas discharges, are quite noisy. Studies of copper vapor arcs have shown that there is a noise voltage superimposed on the dc arc voltage. The frequency spectrum of the noise power density is flat to at least 15 MHz and is detectable at frequencies higher than 8 GHz. Oscillographic voltage traces of this noise show a large number of narrow, *positive* voltage pulses. As the arc current is decreased from 50 to 10 A, the dc arc voltage drops about 3 V. However, the noise amplitude increased dramatically from 4 or 5 V to over 25 V and occasionally pulses

exceeding 60 V or more are seen. It is believed that these positive voltage pulses are produced by sudden decreases in arc current brought about when a cathode spot or perhaps one of its constituent cells decays. The absence of negative voltage pulses in the noise spectrum shows that even though the cathode-spot currents can decrease very quickly from their normal values, they are slow to increase and appear limited by a thermal response time rather than by the arc circuit response time.

As previously discussed, the vacuum arc never dies quietly when the arc current is decreased. With decreasing current a point is reached where only one cathode spot remains. A further decrease in arc current makes this spot very unstable, and while it is carrying a finite current, it suddenly disappears in a time of the order of 10^{-8} sec. This rapid decrease in the current through the circuit inductance will generate a positive voltage surge that may cause the arc to momentarily restrike one or possibly more times before finally extinguishing. This phenomenon is called current chopping and the magnitude of the arc current just prior to arc extinction is called the chopping current. Chopping current is dependent not only on the physical properties of the electrode material and the external circuit, as previously discussed, but also on the rate at which the arc current is decreasing prior to extinction.

III. APPLICATIONS OF THE VACUUM ARC

A. High-Power Vacuum Interrupter

Notwithstanding its complexity and elusiveness to quantitative analysis, the vacuum arc has been harnessed to serve a number of useful purposes. A modern example is the high-power vacuum circuit breaker.

The simplicity and elegance of reliably interrupting large alternating currents in high-voltage circuits by separating two metal contacts enclosed in a vacuum (Fig. 6) had long fascinated scientists and engineers, but early attempts to do so were doomed to failure because of the lack of supporting technologies in vacuum and metallurgical processing. In the early 1920s at the California Institute of Technology, R. A. Millikan, in his research on the field emission of electrons from metals, observed that vacuum gaps had a very high dielectric strength and that many tens of kilovolts would not break down a vacuum gap of only a few millimeters in length. R. W. Sorensen of the same institution applied this phenomenon in his invention of the first vacuum switch. The early work of Professor Sorensen and the General Electric Company showed great promise, but it soon became evident that the sealed vacuum switch of the early 1930s could not offer the high reliability demanded by the electric utilities, for the reasons already

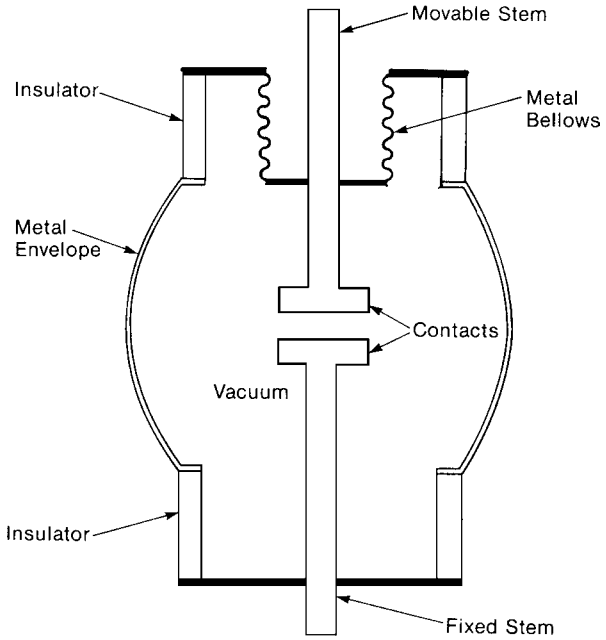


FIGURE 6 The basic elements of a vacuum switch.

mentioned. Twenty years later, after substantial progress had been made in vacuum and metallurgical processing, the General Electric Company took a fresh look at this old problem and successfully developed a high-power vacuum interrupter. The potential advantages of current interruption in a vacuum can be more easily understood by a brief review of the operation of a vacuum switch in a high-voltage, ac circuit.

When the simplified vacuum switch shown in Fig. 6 is closed with the two electrodes in contact and normal current flowing, some heat is developed. Since there are no convection losses in a vacuum and since radiation losses are negligible, the heat generated at the contacts must be carried out by conduction along the leads. Therefore, contact resistance must be low to avoid excessive temperature rise.

When a fault current develops, for example, from a short circuit, an actuating mechanism quickly separates the contacts and a vacuum arc forms between the contacts. Normally, the separation of the contacts does not exceed 0.5 in., and were it not for the fact that alternating current is involved, the high-current arc would continue indefinitely. However, as the first zero is approached on the ac wave, the arc becomes unstable and extinguishes in times of the order of 10^{-8} sec. As discussed in Section II.E, an important characteristic of the vacuum arc is that it does not extinguish at the normal current zero, but at a finite current before current zero, as shown in Fig. 7. The current at which this occurs is called the chopping current. For a short vacuum arc, this current is dependent on the

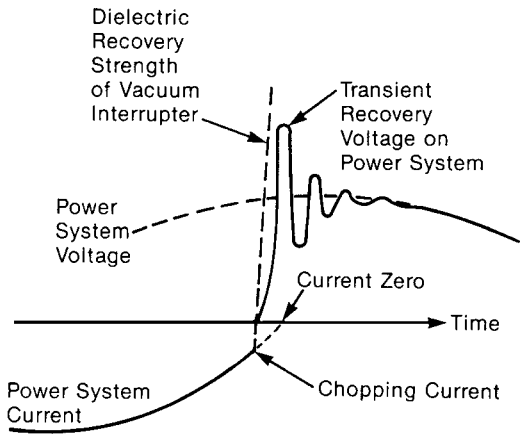


FIGURE 7 Interruption of a fault current by a vacuum circuit breaker and the subsequent transient recovery voltage.

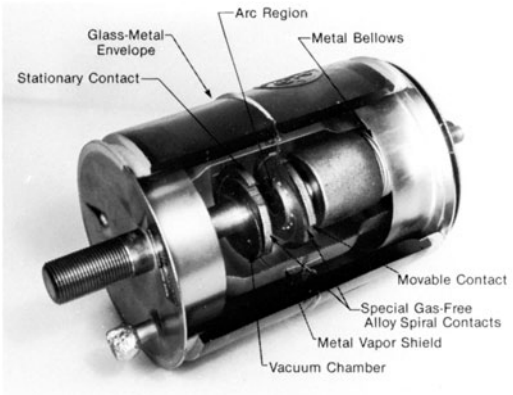


FIGURE 8 A modern 500-MVA vacuum interrupter.

physical characteristics of the contact material (Section II.E). This abrupt cessation of arc current can be a source of difficulty in inductive circuits because of the insulation damage produced by the high-voltage surges. The vacuum switch is particularly susceptible to this difficulty, because the recovery of the electric strength across the gap is so rapid as to permit the development of high overvoltages in the apparatus unless the chopping current is limited to a sufficiently low level.

After current zero, the voltage across the switch is reversed in polarity and—depending on transient circuit conditions—may build up at the rate of 5 to 10 kV/μsec or more. If the rate of dielectric recovery strength of the vacuum switch gap exceeds the rate of rise of the impressed voltage, as shown in Fig. 7, the arc will not restrike and the circuit will have been interrupted. The vacuum switch is unique in that the conducting medium necessary to support the arc is supplied solely by the erosion of the contacts while arcing. When the arc is extinguished, the rapid rate of dispersion and condensation of the metal vapor in the gap determines in part the fast recovery characteristics of the switch. After extinction of the arc, the residual evaporation from the cathode will be negligibly small. At the anode, on the other hand, where melting at the anode spot may have occurred at high current, the residual evaporation may be considerable and have a pronounced effect on the recovery strength. It is clear that anode spots should be avoided if possible.

The advantages to be gained by using vacuum switches for current interruption are many. Because the high breakdown strength makes possible a vacuum gap less that 0.5 in. in length, the actuating mechanism for opening the switch can be relatively simple and fast acting. In the vacuum switch, current interruption occurs at the first current zero after opening; thus the arcing time is usually less than one-half cycle and the energy dissipated in the switch is comparatively small. The fast recovery of the vacuum gap eliminates the necessity for an interrupting medium, such as oil or gas, with its attendant maintenance problems. With a completely sealed switch, there is no fire or explosion hazard.

Figure 8 shows a modern vacuum interrupter capable of interrupting a fault current of over 30,000 A at 15.5 kV. Although rather simple in appearance, the vacuum interrupter contains a considerable amount of sophisticated technology. The device is contained in an evacuated glass-metal envelope. One electrical contact is fixed in position and the other is attached to the vacuum envelope through a flexible metal belows so that it can be moved to open and close the switch. The contacts are surrounded by a metal shield to prevent the metal vapor from the arc from condensing on the glass envelope and spoiling its insulating properties.

The electrical contacts are the most important element in the vacuum switch and are worthy of a detailed description. The ideal material for vacuum switch contacts must satisfy several requirements simultaneously. Foremost, the material must be gas free. If the contacts contain gas it will be released on arcing and accumulate in the switch on successive operations, thus destroying the vacuum and causing the switch to break down at low voltages when the contacts are open. The switch must have a rapid recovery of electric strength immediately after arcing and a high ultimate breakdown strength. The contacts must not weld while carrying high momentary currents or when closing in on a short circuit. This antiwelding characteristic requires that the electrical resistance of the contacts be low, which will also minimize heating during the flow of normal continuous current. Finally, the arc drawn between the contacts must be stable at low currents to prevent the generation of overvoltages by current chopping.

Since each of these characteristics requires, in general, the exploitation of different physical properties of the

contact material, these demands are not easily satisfied simultaneously and, in some instances, may even be contradictory. For example, zinc, which has a very high vapor pressure, has a chopping current of only 0.5 A. However, the recovery strength is very poor because of the high rate at which metal vapor continues to pour off the anode after current zero. Medium-vapor-pressure metals such as copper, which has a good electrical conductivity, are generally soft and tend to stick or weld easily in a vacuum. They are also incapable of maintaining their shape under the high mechanical stresses of rapid opening and reclosing normally encountered in an interrupter. Low-vapor-pressure refractory metals such as tungsten are hard and do not weld, but have a severe chopping problem. It is apparent then that since no ideal contact material can be found if one is limited to pure metals, composite materials must be considered.

The ideal composite material, then, should be hard and have a relatively high melting point, but also be good electrical conductor and contain at least one component with a high vapor pressure. The vacuum interrupter shown in Fig. 8 has contacts made of a two-phase binary alloy of cooper with a few percent of bismuth. In the liquid phase the bismuth is soluble in copper, but on solidification it precipitates out in the grain boundaries and on the surface of the copper. Since the bismuth is virtually insoluble in the copper, the high electrical and thermal conductivity of the copper is retained. The bismuth precipitate in the copper grain boundaries hardens the alloy and produces a weak, brittle weld interface that is easily broken on impact by the vacuum interrupter opening mechanism. Finally, the presence of the high-vapor-pressure bismuth during arcing reduces current chopping. The copper and bismuth are made remarkably free of gas by zone refining, a technique developed by the semiconductor industry for purifying silicon.

As previously mentioned (Section II.C), molten anode spots tend to form on the positive electrode during high-current arcing. Their formation should be avoided in a vacuum switch because of the adverse effect on the recovery strength of the vacuum gap immediately after current zero when the voltage is building up rapidly in the reverse direction. A hot anode spot will not only continue to evaporate metal vapor into the vacuum gap and reduce its recovery strength, but will also favor the formation of cathode spots as it becomes the new cathode. The formation of anode spots has been minimized in the interrupter shown in Fig. 8 by the introduction of spirals in the outer section of the disk contacts. Current flowing in the spiral portions of the contacts produces a magnetic field that exerts a force on the arc column and keeps it moving, thus distributing the arc energy over a large portion of the positive contact.

B. Triggered Vacuum Gap

The triggered vacuum gap (TVG) is basically a vacuum switch with permanently separated contacts and a third electrode to which an impulse voltage may be applied to rapidly initiate a vacuum arc between the main contacts. The advantages of such a device are the high current-carrying capacity and short breakdown time. Triggered voltage gaps can be designed to break down in less than 100 nsec and carry currents of at least 10^6 A. They have found applications as high-speed protective devices to short out components in danger of damage by overvoltage or overcurrent. They are also used in commutating circuits and as "make switches" in fusion devices.

The design and vacuum processing of a TVG is similar to that of a vacuum switch. The selection of a proper gas-free electrode material for a given gap will depend on the application requirements. However, one need not be concerned about contact welding and resistance to mechanical deformation since the electrodes are permanently separated and do not move. Relaxation of these requirements gives more freedom in the selection of electrode geometries to carry high currents and of electrode materials for rapid breakdown, fast recovery, and high ultimate breakdown strength if any of these features is required.

As discussed in Section II.A, the presence of both electrons and ions is required in a high-voltage vacuum gap to initiate breakdown and establish a vacuum arc. These can be conveniently injected into the vacuum gap in the form of a plasma to give rapid breakdown with a minimum of jitter. One such arrangement for plasma injection is shown in Fig. 9. A ceramic cylinder is coated with titanium hydride only a few millimeters thick. A V-notched groove is then cut through the metallic hydride into the ceramic, forming a ceramic gap. A shield cap is placed on one end of the ceramic cylinder, and a lead wire is attached to the cap. This assembly is inserted into a conical recess in the cathode electrode of the main vacuum gap as shown.

To operate the trigger, a positive voltage pulse is applied to the trigger lead. The ceramic gap breaks down and an arc is established between the titanium hydride electrodes, thus releasing hydrogen and titanium vapor, which are ionized and sustain the discharge. Expansion and magnetic forces produced by the discharge current loop drive the plasma out of the conical recess into the main gap. As the plasma spreads out into the main gap, a high-current glow discharge is first established between the main electrodes. This glow is rapidly transformed into a high-current vacuum arc. It would appear that the main gap discharge uses the cathode spots already established on the trigger electrode in the initial stages of buildup. Measurements indicate that, with peak current pulses of 10 A through the trigger electrode, the main gap will break

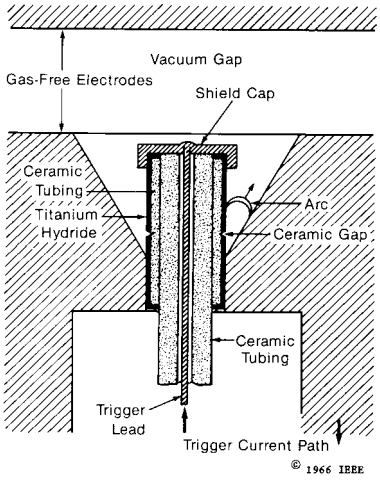


FIGURE 9 Cross-sectional view of an arrangement for triggering a vacuum gap. [Reproduced with permission from Lafferty, J. M. (1966). *Proc. IEEE* **54,** 23. © 1966 IEEE.]

down in less than 0.1 μsec with jitter times of about 30 nsec when 30 kV is applied to the main gap with an electrode separation of $\frac{1}{8}$ in. The trigger energy required is less than 0.01 J. The main gap may be broken down with trigger pulses as low as 50 V; however, longer delay times result.

The quantity of hydrogen released on firing the trigger is extremely minute and there is no accumulative pressure rise due to hydrogen. The presence of the hydride is by no means essential to the operation of the trigger. Breakdown of the gap can be produced by ionization of metal vapor eroded from the trigger electrodes. The energy required to produce breakdown of the gap with an unloaded trigger is about 10 times that required when hydrided.

Reversal of the trigger pulse polarity or placement of the trigger in the anode electrode of the main gap requires more energy for triggering. The principal reason for this is that the cathode spots for the trigger discharge are no longer on the cathode of the main gap and cannot share in the development of the main discharge.

Other trigger electrode configurations are possible for injecting plasma into the main vacuum gap. For example, a coaxial plasma gun may be used that can be well shielded from the main gap discharge. However, such devices all require more trigger energy than the device shown in Fig. 9.

A number of triggered vacuum gaps have been developed over a wide range of currents and voltages for various applications. Some fast TVGs can carry only capacitor discharge current for a few microseconds, while others can carry 60-Hz power line currents of tens of thousands of amperes for one-half cycle. The operating voltages may range from a few hundred volts to 100 kV. Figure 10 shows a TVG with a rating of 73 kV and 125,000 A that is used to protect the series-compensating capacitors on the

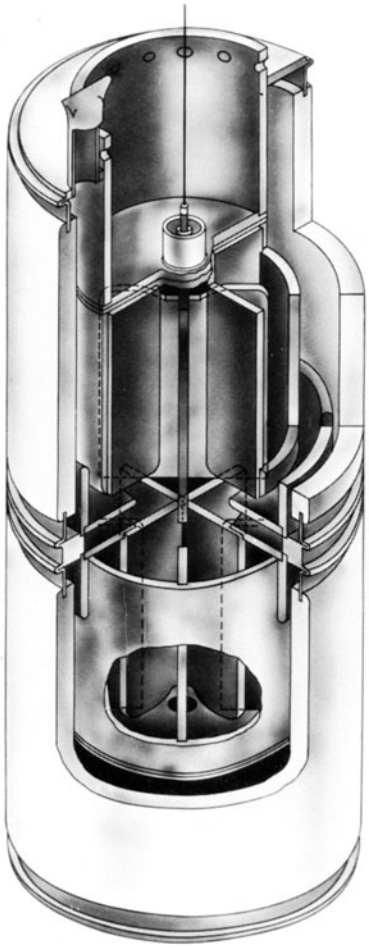


FIGURE 10 Sketch of a 9000-MVA triggered vacuum gap used to protect the series-compensating capacitors on the 500-kV. Pacific intertie ac transmission line. Two communicating gaps in series with an interdigital paddle-wheel geometry were used to disperse the vacuum arc and avoid anode-spot formation.

500-kV Pacific intertie ac transmission line that runs from the Columbia River to Los Angeles in the United States.

Triggered vacuum gaps have also been designed with a movable contact so that they can be made to conduct quickly with the contacts open and then closed to protect the gap if a large current is to flow for an appreciable time. This can also be accomplished by placing a TVG and vacuum switch in parallel.

C. Vacuum Fuse

The vacuum fuse is another example of the use of high-current vacuum arcs for current interruption and the protection of power circuits. Its design is similar to that of a triggered vacuum gap with the fixed arcing electrodes bridged by a fuse element that conducts current under normal conditions. On the incident of a fault or other abnormal

condition that produces a large current flow, the fuse element melts and opens the circuit between the electrodes. This initiates a vacuum arc in the same way that it is established when the contacts of a vacuum switch are separated while carrying current, as described in Section II.A. Once the vacuum arc is established, the device functions like a vacuum interrupter with its contacts in the open position.

hThe relatively short length of the fuse element allows the heat generated in it by the normal load current to be conducted out through the heavy leads supporting the arcing electrodes. The arcing electrodes also act as a heat sink to conduct heat away from the fuse element under temporary overload conditions. Under fault conditions the heat cannot escape quickly enough, and the fuse melts at a necked down portion in the center of the bridging element and creates a vacuum arc. The arc quickly melts and vaporizes the remainder of the fuse element, creating conditions essentially identical to those in an open vacuum interrupter. Obviously the fuse bridge element must be made from gas-free material or the fuse will not function properly in a high-voltage ac circuit.

SEE ALSO THE FOLLOWING ARTICLES

DIELECTRIC GASES • PLASMA SCIENCE AND ENGINEERING • POWER TRANSMISSION, HIGH VOLTAGE • PULSED POWER SYSTEMS • VACUUM TECHNOLOGY

BIBLIOGRAPHY

Boxman, R., Martin, P., and Sanders, D., eds. (1995). "Handbook of Vacuum Arc Science and Technology," Noyes Publications, Park Ridge, NJ.

Kimblin, C. W. (1969). Anode voltage drop and anode spot formation in dc vacuum arcs. *J. Appl. Phys*. **40,** 1744–1752.

Kimblin, C. W. (1971). Vacuum arc ion currents and electrode phenomena. *Proc. IEEE* **59,** 546–555.

Lafferty, J. M., ed. (1980). "Vacuum Arcs" [with contributions by James D. Cobine, Günter Ecker, George A. Farrall, Allan N. Greenwood, and L. P. Harris]. Wiley, New York.

Miller, H. C. (1984). A review of anode phenomena in vacuum arcs. *Proc. Int. Symp. Discharges Electric, Insulation Vacuum 11th, Berlin, 1984,* **1,** 115–121.

Reese, M. P. (1973). A review of the development of the vacuum interrupter. *Philos. Trans. R. Soc. London A* **275,** 121–129.