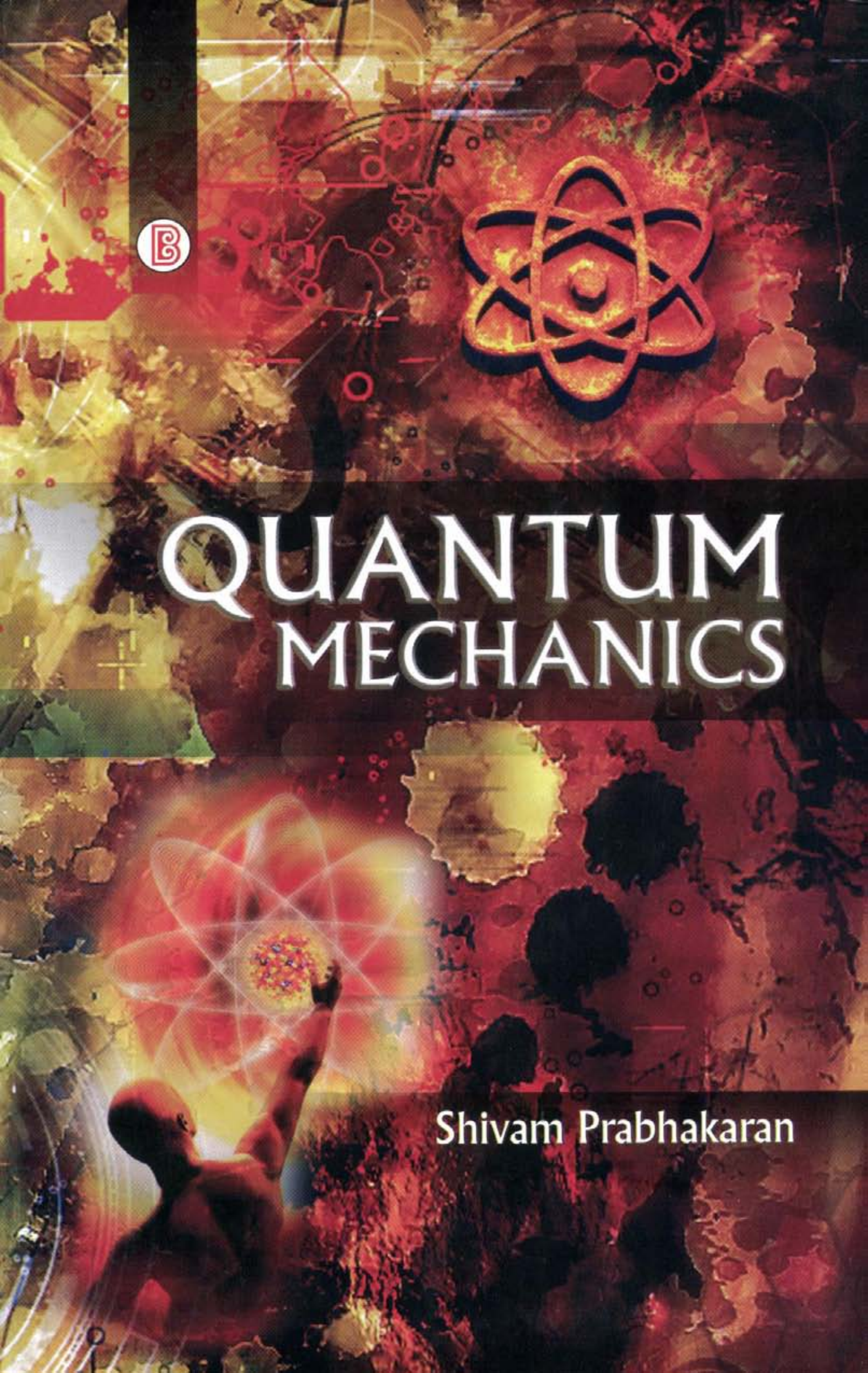


QUANTUM MECHANICS

Shivam Prabhakaran

Quantum Mechanics

Shivam Prabhakaran

Book Enclave
Jaipur India

First Published : 2009
ISBN : 978-81-8152-234-4

Published by
Book Enclave
Jain Bhawan, Opp. N.E.I., Shanti Nagar, Jaipur - 302006 Tel. 0141-2221456
Showroom
G-13, S.S. Tower, Dhamani Street, Chaura Rasta, Jaipur
Tel. 0141-2311900, Fax: +91-141-2311900,
E-mail: bookenclave@yahoo.com • bookenclave@gmail.com
Website : www.bookenclave.com
Laser Typesetting by
Shubham Computers, Jaipur
Printed at
Roshan Offset Printers, Delhi

Preface

Quantum Mechanics is a theory of Mechanics, a branch of Physics that deals with the Motion of bodies and associated physical quantities such as Energy and Momentum. Quantum Mechanics has had enormous success in explaining many of the features of our world. The individual behaviour of the Microscopic Particles that make up all forms of matter can often only be satisfactorily described using Quantum Mechanics.

Quantum mechanics is important for understanding how individual atoms combine to form chemicals. It provides quantitative insight into chemical bonding processes by explicitly showing which molecules are energetically favourable to which others, and by approximately how much. This book is intended to provide a comprehensive coverage of the major aspects of quantum mechanics. The most likely audience for the book consists of students and teachers of modern physics, mechanics and engineering.

Shivam Prabhakaran

"This page is Intentionally Left Blank"

Contents

"This page is Intentionally Left Blank"

1

Introduction to Quantum Physics

For centuries, man has wondered on phenomena and processes happening around him. As time passed, he was successful in applying his intuition and common sense in comprehending the stars, galaxies and their behaviour, but they fail in the microscopic world of molecules, atoms and sub-atomic particles.

Quantum theory provides us with the rules and regulations of the miniature world. These rules are phenomenally successful in accounting for the properties of atoms, molecules, and their constituents, and form the basis of understanding the fundamental properties of all matter. In fact, one may say that the greatest success story of the 20th-century physics is to confirm that this theory works, without a single exception, in spite of critical examination by some of the best minds spanning decades of time.

The conceptual foundation of quantum theory is mysterious. It led to intense debates among scientists, and confused many. Niels Bohr, one of the most prominent scientists in this domain, once remarked, "You have not studied quantum mechanics well if you aren't confused by it." Albert Einstein, the greatest physicist of the 20th century, never approved of this theory. Bizarre though it may seem, quantum physics has led physicists step by step to a deeper view of the reality, and has answered many fundamental questions.

Quantum physics is a branch of science that deals with discrete, indivisible units of energy called quanta as described by the Quantum Theory. There are five main ideas represented in Quantum Theory:

1. Energy is not continuous, but comes in small but discrete units.

2. The elementary particles behave both like particles and like waves.
3. The movement of these particles is inherently random.
4. It is *physically impossible* to know both the position and the momentum of a particle at the same time.
5. The atomic world is nothing like the world we live in.

While at a glance this may seem like just another strange theory, it contains many clues as to the fundamental nature of the universe and is more important then even relativity in the grand scheme of things (if any one thing at that level could be said to be more important then anything else). Furthermore, it describes the nature of the universe as being much different then the world we see. As Niels Bohr said, "Anyone who is not shocked by quantum theory has not understood it."

Particle / Wave Duality

Particle/wave duality is perhaps the easiest way to get aquatinted with quantum theory because it shows, in a few simple experiments, how different the atomic world is from our world.

First let's set up a generic situation to avoid repetition. In the centre of the experiment is a wall with two slits in it. To the right we have a detector. What exactly the detector is varies from experiment to experiment, but it's purpose stays the same: detect how many of whatever we are sending through the experiment reaches each point. To the left of the wall we have the originating point of whatever it is we are going to send through the experiment. That's the experiment: send something through two slits and see what happens. For simplicity, assume that nothing bounces off of the walls in funny patterns to mess up the experiment.

First try the experiment with bullets. Place a gun at the originating point and use a sandbar as the detector. First try covering one slit and see what happens. You get more bullets near the centre of the slit and less as you get further away. When you cover the other slit, you see the same thing with respect to the other slit. Now open both slits. You get the sum of the result of opening each slit. The most bullets are found in the middle of the two slits with less being found the further you get from the centre.

Well, that was fun. Let's try it on something more interesting: water waves. Place a wave generator at the originating point and detect using a wave detector that measures the height of the waves that pass. Try it with one slit closed. You see a result just like that of the bullets. With the other slit closed the result is the same. Now try it with both slits open. Instead of getting the sum of the results of each slit being open, you see a wavy pattern; in the centre there is a wave greater then the sum of what appeared there each time only one slit was open. Next to that large wave was a wave much smaller then what appeared there during either of the two single slit runs. Then the pattern repeats; large wave, though not nearly as large as the centre one, then small wave. This makes sense; in some places the waves reinforced each other creating a larger wave, in other places they canceled out. In the centre there was the most overlap, and therefore the largest wave. In mathematical terms, instead of the resulting intensity being the sum of the squares of the heights of the waves, it is the square of the sum.

While the result was different from the bullets, there is still nothing unusual about it; everyone has seen this effect when the waves from two stones that are dropped into a lake in different places overlap. The difference between this experiment and the previous one is easily explained by saying that while the bullets each went through only one slit, the waves each went through both slits and were thus able to interfere with themselves.

Now try the experiment with electrons. Recall that electrons are negatively charged *particles* that make up the outer layers of the atom. Certainly they could only go through one slit at a time, so their pattern should look like that of the bullets, right? Let's find out. Place an electron gun at the originating point and an electron detector in the detector place. First try opening only one slit, then just the other. The results are just like those of the bullets and the waves. Now open both slits. *The result is just like the waves.*

There must be some explanation. After all, an electron couldn't go through both slits. Instead of a continuous stream of electrons, let's turn the electron gun down so that at any one time only one electron is in the experiment. Now the electrons won't be able to cause trouble since there is no one else to interfere with. The result should now look like the bullets. But it doesn't! It would seem that the electrons do go through both slits.

This is indeed a strange occurrence; we should watch them ourselves to make sure that this is indeed what is happening. So, we put a light behind the wall so that we can see a flash from the slit that the electron went through, or a flash from both slits if it went through both. Try the experiment again. As each electron passes through, there is a flash in only one of the two slits.

Obviously the light is causing problems. Perhaps if we turned down the intensity of the light, we would be able to see them without disturbing them. When we try this, we notice first that the flashes we see are the same size. Also, some electrons now get by without being detected. This is because light is not continuous but made up of particles called photons. Turning down the intensity only lowers the number of photons given out by the light source The particles that flash in one slit or the other behave like the bullets, while those that go undetected behave like waves.

Well, we are not about to be outsmarted by an electron, so instead of lowering the intensity of the light, why don't we lower the frequency. The lower the frequency the less the electron will be disturbed, so we can finally see what is actually going on. Lower the frequency slightly and try the experiment again. We see the bullet curve. After lowering it for a while, we finally see a curve that looks somewhat like that of the waves! There is one problem, though. Lowering the frequency of light is the same as increasing it's wavelength, and by the time the frequency of the light is low enough to detect the wave pattern the wavelength is longer then the distance between the slits so we can no longer see which slit the electron went through.

So have the electrons outsmarted us? Perhaps, but they have also taught us one of the most fundamental lessons in quantum physics - an observation is only valid in the context of the experiment in which it was performed. If you want to say that something behaves a certain way or even exists, you must give the context of this behaviour or existence since in another context it may behave differently or not exist at all. We can't just say that an electron is a particle, since we have already seen proof that this is not always the case. We can only say that when we observe the electron in the two slit experiment it behaves like a particle. To see how it would behave under different conditions, we must perform a different experiment.

The Copenhagen Interpretation

So sometimes a particle acts like a particle and other times it acts like a wavé. So which is it? According to Niels Bohr, who worked in Copenhagen when he presented what is now known as the Copenhagen interpretation of quantum theory, the particle is what you measure it to be. When it looks like a particle, it is a particle. When it looks like a wave, it is a wave. Furthermore, it is meaningless to ascribe any properties or even existence to anything that has not been measured. Bohr is basically saying that nothing is real unless it is observed.

While there are many other interpretations of quantum physics, all based on the Copenhagen interpretation, the Copenhagen interpretation is by far the most widely used because it provides a 'generic' interpretation that does not try to say any more then can be proven. Even so, the Copenhagen interpretation does have a flaw that we will discuss later. Still, since after 70 years no one has been able to come up with an interpretation that works better then the Copenhagen interpretation, that is the one we will use. We will discuss one of the alternatives later.

The Wave Function

In 1926, just weeks after several other physicists had published equations describing quantum physics in terms of matrices, Erwin Schrodinger created quantum equations based on wave mathematics, a mathematical system that corresponds to the world we know much more then the matrices. After the initial shock, first Schrodinger himself then others proved that the equations were mathematically equivalent. Bohr then invited Schrodinger to Copenhagen where they found that Schrodinger's waves were in fact nothing like real waves. For one thing, each particle that was being described as a wave required three dimensions. Even worse, from Schrodinger's point of view, particles still jumped from one quantum state to another; even expressed in terms of waves space was still not continuous. Upon discovering this, Schrodinger remarked to Bohr that "Had I known that we were not going to get rid of this damned quantum jumping, I never would have involved myself in this business."

Unfortunately, even today people try to imagine the atomic world as being a bunch of classical waves. As Schrodinger found out, this could not be farther from the truth. The atomic world is nothing like our world, no matter how much we try to pretend it is. In many ways, the success of Schrodinger's equations has prevented people from thinking more deeply about the true nature of the atomic world.

The Collapse of the Wave Function

So why bring up the wave function at all if it hampers full appreciation of the atomic world? For one thing, the equations are much more familiar to physicists, so Schrodinger's equations are used much more often than the others. Also, it turns out that Bohr liked the idea and used it in his Copenhagen interpretation. Remember the experiment with electrons? Each possible route that the electron could take, called a ghost, could be described by a wave function. As we shall see later, the 'damned quantum jumping' insures that there are only a finite, though large, number of possible routes. When no one is watching, the electron take every possible route and therefore interferes with itself. However, when the electron is observed, it is forced to choose one path. Bohr called this the "collapse of the wave function". The probability that a certain path will be chosen when the wave function collapses is essentially the square of the path's wave function.

Bohr reasoned that nature likes to keep its possibilities open, and therefore follows every possible path. Only when observed is nature forced to choose only one path, so only then is just one path taken.

The Uncertainty Principle

If we are going to destroy the wave pattern by observing the experiment, then we should at least be able to determine exactly where the electron goes. Newton figured that much out back in the early eighteenth century; just observe the position and momentum of the electron as it leaves the electron gun and we can determine exactly where it goes.

Well, fine. But how exactly are we to determine the position and the momentum of the electron? If we disturb the electrons just in seeing if they are there or not, how are we possibly going to determine both their position and momentum? Still, a clever enough person, say Albert Einstein, should be able to come up with something, right?

Unfortunately not. Einstein did actually spend a good deal of his life trying to do just that and failed. Furthermore, it turns out that if it were possible to determine both the position and the momentum at the same time, Quantum Physics would collapse. Because of the latter, Werner Heisenberg proposed in 1925 that it is in fact *physically impossible* to do so. As he stated it in what now is called the Heisenberg Uncertainty Principle, if you determine an object's position with uncertainty x, there must be an uncertainty in momentum p, such that xp > *h*/4pi, where *h* is Planck's constant (which we will discuss shortly). In other words, you can determine *either* the position *or* the momentum of an object as accurately as you like, but the act of doing so makes your measurement of the other property that much less. Human beings may someday build a device capable of transporting objects across the galaxy, but no one will *ever* be able to measure both the momentum and the position of an object at the same time. This applies not only to electrons but also to objects such as tennis balls and toasters, though for these objects the amount of uncertainty is so small compared to there size that it can safely be ignored under most circumstances.

The EPR Experiment

"God does not play dice" was Albert Einstein's reply to the Uncertainty Principle. Thus being his belief, he spent a good deal of his life after 1925 trying to determine both the position and the momentum of a particle. In 1935, Einstein and two other physicists, Podolski and Rosen, presented what is now known as the EPR paper in which they suggested a way to do just that. The idea is this: set up an interaction such that two particles are go off in opposite directions and do not interact with anything else. Wait until they are far apart, then measure the momentum of one and the position of the other. Because of conservation of momentum, you can determine the momentum of the particle not measured, so when you measure its position you know both its momentum and position. The only way quantum physics could be true is if the particles could communicate faster than the speed of light, which Einstein reasoned would be impossible because of his Theory of Relativity.

In 1982, Alain Aspect, a French physicist, carried out the EPR experiment. He found that *even if information needed to be* *communicated faster than light to prevent it, it was not possible to determine both the position and the momentum of a particle at the same time.* This does not mean that it is possible to send a message faster than light, since viewing either one of the two particles gives no information about the other. It is only when both are seen that we find that quantum physics has agreed with the experiment. So does this mean relativity is wrong? No, it just means that the particles do not communicate by any means we know about. All we know is that every particle knows what every other particle it has ever interacted with is doing.

The Quantum and Planck's Constant

So what is that *h* that was so importantce in the Uncertainty Principle? Well, technically speaking, its 6.63 × 10^{-34} joule-seconds. It's call Planck's constant after Max Planck who, in 1900, introduced it in the equation $E=hv$ where E is the energy of each quantum of radiation and v is its frequency. What this says is that energy is not continuous as everyone had assumed but only comes in certain finite sizes based on Planck's constant.

At first physicists thought that this was just a neat mathematical trick Planck used to explain experimental results that did not agree with classical physics. Then, in 1904, Einstein used this idea to explain certain properties of light—he said that light was in fact a particle with energy $E=hv$. After that the idea that energy isn't continuous was taken as a fact of nature—and with amazing results. There was now a reason why electrons were only found in certain energy levels around the nucleus of an atom. Ironically, Einstein gave quantum theory the push it needed to become the valid theory it is today, though he would spend the rest of his life trying to prove that it was not a true description of nature.

Also, by combining Planck's constant, the constant of gravity, and the speed of light, it is possible to create a quantum of length (about 10^{-35} metre) and a quantum of time (about 10^{-43} sec), called, respectively, Planck's length and Planck's time. While saying that energy is not continuous might not be too startling to the average person, since what we commonly think of as energy is not all that well defined anyway, it is startling to say that there are quantities of space and time that cannot be broken up into smaller pieces. Yet it is exactly this that gives nature a finite number of routes to take when an electron interferes with itself.

Although it may seem like the idea that energy is quantized is a minor part of quantum physics when compared with ghost electrons and the uncertainty principle, it really is a fundamental statement about nature that caused everything else we've talked about to be discovered. And it is always true. In the strange world of the atom, anything that can be taken for granted is a major step towards an 'atomic worldview'.

Schrodinger's Cat

There was a problem with the Copenhagen interpretation? Well, you now know enough of what quantum physics *is* to be able to discuss what it *isn't*, and by far the biggest thing it isn't is complete. Sure, the math seems to be complete, but the theory includes absolutely nothing that would tie the math to any physical reality we could imagine. Furthermore, quantum physics leaves us with a rather large open question: *what is reality?* The Copenhagen interpretation attempts to solve this problem by saying that reality is what is measured. However, the measuring device itself is then not real until *it* is measured. The problem, which is known as the measurement problem, is when does the cycle stop?

Remember that when we last left Schrodinger he was muttering about the 'damned quantum jumping.' He never did get used to quantum physics, but, unlike Einstein, he was able to come up with a very real demonstration of just how incomplete the physical view of our world given by quantum physics really is. Imagine a box in which there is a radioactive source, a Geiger counter (or anything that records the presence of radioactive particles), a bottle of cyanide, and a cat. The detector is turned on for just long enough that there is a fifty-fifty chance that the radioactive material will decay. If the material does decay, the Geiger counter detects the particle and crushes the bottle of cyanide, killing the cat. If the material does not decay, the cat lives. To us outside the box, the time of detection is when the box is open. At that point, the wave function collapses and the cat either dies or lives. However, until the box is opened, the cat is both dead and alive.

On one hand, the cat itself could be considered the detector; its presence is enough to collapse the wave function. But in that case, would the presence of a rat be enough? Or an amoeba? Where is the line drawn? On the other hand, what if you replace the cat with a human (named 'Wigner's friend' after Eugene Wigner, the physicist who developed many derivations of the Schrodinger's cat experiment). The human is certainly able to collapse the wave function, yet to us outside the box the measurement is not taken until the box is opened. If we try to develop some sort of 'quantum relativity' where each individual has his own view of the world, then what is to prevent the world from getting "out of sync" between observers?

While there are many different interpretations that solve the problem of Schrodinger's Cat, one of which we will discuss shortly, none of them are satisfactory enough to have convinced a majority of physicists that the consequences of these interpretations are better than the half dead cat. Furthermore, while these interpretations do prevent a half dead cat, they do not solve the underlying measurement problem. Until a better intrepretation surfaces, we are left with the Copenhagen interpretation and its half dead cat. We can certainly understand how Schrodinger feels when he says, "I don't like it, and I'm sorry I ever had anything to do with it."

The Infinity Problem

There is one last problem that we will discuss before moving on to the alternative interpretation. Unlike the others, this problem lies primarily in the mathematics of a certain part of quantum physics called quantum electrodynamics, or QED. This branch of quantum physics explains the electromagnetic interaction in quantum terms. The problem is, when you add the interaction particles and try to solve Schrodinger's wave equation, you get an electron with infinite mass, infinite energy, and infinite charge. There is no way to get rid of the infinities using valid mathematics, so, the theorists simply divide infinity by infinity and get whatever result the guys in the lab say the mass, energy, and charge should be. Even fudging the math, the other results of QED are so powerful that most physicists ignore the infinities and use the theory anyway. As Paul Dirac, who was one of the physicists who published quantum equations before Schrodinger, said, "Sensible mathematics involves neglecting a quantity when it turns out to be small–not neglecting it just because it is infinitely great and you do not want it!".

Many Worlds

One other interpretation, presented first by Hugh Everett III in 1957, is the many worlds or branching universe interpretation. In this theory, whenever a measurement takes place, the entire universe divides as many times as there are possible outcomes of the measurement. All universes are identical except for the outcome of that measurement. Unlike the science fiction view of 'parallel universes', it is not possible for any of these worlds to interact with each other.

While this creates an unthinkable number of different worlds, it does solve the problem of Schrodinger's cat. Insıead of one cat, we now have two; one is dead, the other alive. However, it has still not solved the measurement problem. If the universe split every time there was more than one possibility, then we would not see the interference pattern in the electron experiment. So when does it split? No alternative interpretation has yet answered this question in a satisfactory way.

Classical Physics from Newton to Einstein

The Scientific Method

The scientific method has four major components:

1. The assumption of an external, objective reality that can be observed.
2. Quantitative experiments on the external objective reality in order to determine its observable properties, and the use of induction to discover its general principles.
3. Validation of the results of these measurements by widespread communication and publication so that other scientists are able to verify them independently. Although scientists throughout history have communicated and published their results, the first scientist to articulate the need for publishing the details of his experimental methods so that other scientists could repeat his measurements was English chemist Robert Boyle, who was strongly influenced by the views of Bacon.
4. Intuiting and formulating the mathematical laws that describe the external objective reality. The most universal laws are those of physics, the most fundamental science. English natural philosopher Isaac Newton was the first scientist to formulate laws that were considered to apply universally to all physical systems.

The last three of these components were all developed in the remarkably brief period from 1620 to 1687, and all by Englishmen!

Newton's Laws and Determinism

In order to understand quantum physics, we must first understand classical physics so that we can see the differences between them.

There are two fundamental assumptions in classical physics. The first fundamental assumption is that the objective world exists independently of any observations that are made on it. To use a popular analogy, a tree falling in the forest produces a sound whether or not it is heard by anyone. While it is possible that observations of the objective world can affect it, its independence guarantees that they do not necessarily affect it.

The second fundamental assumption of classical physics is that both the position and velocity of an object can be measured with no limits on their precision except for those of the measuring instruments. In other words, the objective world is a precise world with no intrinsic uncertainty in it. As we shall see later, quantum theory abandons both of these fundamental assumptions.

Isaac Newton was the first important scientist both to do fundamental experiments and to devise comprehensive mathematical theories to explain them. He invented a theory of gravity to explain the laws of German astronomer and mathematician Johannes Kepler which describe the planetary orbits, made use of the famous free-fall experiments from the leaning tower of Pisa by Italian scientist Galileo Galilei, and invented the calculus in order to give a proper mathematical framework to the laws of motion that he discovered. Newton considered himself to be a natural philosopher, but contemporary custom would accord him the title of physicist. Indeed, he, probably more than any other scientist, established physics as a separate scientific discipline because of his attempts to express his conclusions in terms of universal physical laws.

His three laws of motion can be written as follows:

1. A body moves with constant velocity unless there is a nonzero net force acting on it.
2. The rate of change of the velocity of a body is proportional to the force on the body.

3. If one body exerts a force on another body, the second body exerts an equal and opposite force on the first.

In order to use these laws, the properties of the forces acting on a body must be known. As an example of a force and its properties, Newton's law of gravitation states that the gravitational force between two bodies, such as the earth and the moon, is proportional to the mass of each body and is inversely proportional to the square of the distance between them. This description of the gravitational force, when used together with Newton's second law, explains why the planetary orbits are elliptical. Because of Newton's third law, the force acting on the earth is equal and opposite to the force acting on the moon. Both bodies are constantly changing their speeds and directions because of the gravitational force continually acting on them.

For more than 200 years, after many experiments on every accessible topic of macroscopic nature, Newton's laws came to be regarded by physicists and by much of society as the laws that were obeyed by all phenomena in the physical world. They were successful in explaining all motions, from those of the planets and stars to those of the molecules in a gas. This universal success led to the widespread belief in the principle of determinism, which says that, if the state of a system of objects (even as all-encompassing as the universe) is known precisely at any given time, such as now, the state of the system at any time in the future can in principle be predicted precisely. For complex systems, the actual mathematics might be too complicated, but that did not affect the principle. Ultimately, this principle was thought to apply to living beings as well as to inanimate objects. Such a deterministic world was thought to be completely mechanical, without room for free will, indeed without room for even any small deviation from its ultimate destiny. If there was a God in this world, his role was limited entirely to setting the whole thing into motion at the beginning.

Intrinsic to the principle of determinism was the assumption that the state of a system of objects could be precisely described at all times. This meant, for example, that the position and velocity of each object could be specified exactly, without any uncertainty. Without such exactitude, prediction of future positions and velocities would be impossible. After many, many experiments it seemed clear that only the inevitable imprecision in measuring instruments limited the accuracy of a velocity or position measurement, and nobody doubted that accuracies could improve without limit as measurement techniques improved.

Thermodynamics and Statistical Mechanics, Entropy and the Direction of Time

Thermodynamics is the physics of heat flow and of the interconversion between heat energy and other forms of energy. Statistical mechanics is the theory that describes macroscopic properties such as pressure, volume and temperature of a system in terms of the average properties of its microscopic constituents, the atoms and molecules. Thermodynamics and statistical mechanics are both concerned with predicting the same properties and describing the same processes, thermodynamics from a macroscopic point of view, and statistical mechanics from a microscopic point of view.

In 1850, the German physicist Rudolf Clausius proposed the first law of thermodynamics, which states that energy may be converted from one form to another, such as heat energy into the mechanical rotation of a turbine, but it is always conserved. Since 1905 when German-Swiss-American physicist Albert Einstein invented the special theory of relativity, we know that energy and matter can be converted into each other. Hence, the first law actually applies jointly to both matter and energy. This law is probably the most fundamental one in nature. It applies to all systems, no matter how small or large, simple or complex, whether living or inanimate. We do not think it is ever violated anywhere in the universe. No new physical theory is ever proposed without checking to see whether it upholds this law.

The second law of thermodynamics can be stated in several ways. The first statement of it, made by Rudolf Clausius in 1850, is that heat can flow spontaneously from a hot to a cold object but it cannot spontaneously pass from a cold to a hot object. The second statement of the second law was made later by Scottish physicist William Thomson Kelvin and German physicist Max Planck: Heat energy cannot be completely transformed into mechanical energy, but mechanical energy can be completely transformed into heat energy. The third statement of the second law depends on a new concept, that of entropy.

Entropy is related to the amount of disorder and order in the system. Decreasing entropy is equivalent to decreasing disorder or disorganization (increasing order or organization) of an object or system; while increasing entropy is equivalent to increasing disorder or disorganization.

It turns out that the second law of thermodynamics can be stated in the following way: Natural processes of an isolated macroscopic system normally proceed in the direction of maximum probability (maximum disorder), which is the direction of maximum number of distinguishable arrangements of the system. (It is highly improbable, although not totally impossible, for them to proceed in the opposite direction.) The forward direction of time is the direction in which entropy increases. Thus, the second law of thermodynamics can be restated in terms of entropy: Natural processes of an isolated macroscopic system always proceed in the direction of increasing entropy (disorder).

The direction of time can also be inferred from the first two statements of the second law of thermodynamics: (1) The unidirectional flow of heat from hot to cold bodies, and (2) the possibility of total conversion of mechanical energy to heat energy, but ~~not the reverse~~.

A mistake made by some people is to think that the second law applies to individual objects or systems, such as automobiles, plants, or human bodies, even if they are not isolated from the rest of the universe, and that this is the reason that such objects decay and disintegrate with time. This is a fallacy, however, because the second law does not prevent the entropy of an individual object from continuously decreasing with time and thus becoming more ordered and organized as long as it receives energy from something else in the universe whose entropy continues to increase. In our solar system, it is primarily the sun's entropy that continually increases as its fuel is burned and it becomes more disordered.

An extremely important property of Newton's laws is that they are time reversal invariant. What this obscure-sounding term means is that, if the direction of time is reversed, the directions of motion of all particles are also reversed, and this reversed motion is completely allowed by Newton's laws. In other words, the motion in reversed time is just as valid as the motion in forward time, and nature herself does not distinguish between the two. A simple example of this is the time-reversed motion of a thrown baseball, which follows a parabolic trajectory in either the forward or the reversed direction. Without seeing the act of throwing, and without air resistance, we would not be able to distinguish the forward parabola from the reversed parabola. Another way to state it is that a movie of a thrown baseball seems just as valid to us if it is run in the reverse direction as in the forward direction. Time reversal invariance is also apparent in the seemingly random motion of the molecules in a gas. If we could see their motion in a movie and then reverse it, we could not distinguish between the forward motion and the reversed motion.

However, if we consider the motion of an object containing many ordered particles (for example, with a recognizable size, shape, position, velocity, and orientation), we encounter a different phenomenon. It is easy to tell the difference between the reversed and forward motions of a person, a horse, a growing plant, a cup falling from a table and breaking, and most other examples from everyday life. Another example is the free expansion of a gas that initially is confined to one side of a box by a membrane. If the membrane is broken, the gas immediately expands into the other side (initially assumed to be evacuated), and we can easily tell the time reversed motion from the forward motion. In all of these cases, the motion at the individual molecule level is time reversal invariant, but it is clear that the gross motion of the macroscopic object is not.

Our question now is, "Why does nature seem to be time reversal invariant at the individual, or few, particle level, but apparently not at the level of many particles contained in an ordered system such as any common macroscopic object?" In classical physics, irreversibility is always due to the second law of thermodynamics, which determines the forward direction of time. The forward direction is apparent after the cup has fallen and broken because the broken cup is more disordered (has higher entropy) than the unbroken cup. However, even before the cup breaks, a detailed calculation would show that the entropy of the combined system of cup, gravitational force, and earth increases as the cup falls. The entropy of the system of moving horse or person, gravitational force, earth, and surroundings increases with time because the motion dissipates energy and increases the disorder in the body, earth, and surroundings.

Electromagnetism

French physicist Charles Augustin de Coulomb discovered the force law obeyed by stationary, electrically charged objects between 1785 and 1791. In 1820, Danish physicist Hans Christian Oersted discovered that an electric current produces a magnetic field, and showed that a magnetic field exerted a force on a current-carrying wire. From 1820 to 1827, French physicist Andre Ampere extended these discoveries and developed the mathematical relationship describing the strength of the magnetic field as a function of current. In 1831, English chemist and physicist Michael Faraday discovered that a changing magnetic field, which he explained in terms of changing magnetic lines of force, produces an electric current in a wire. This was a giant step forward because it was the forerunner of the concept of force fields, which are used to explain all forces in nature today.

These disparate phenomena and theories were all pulled together into one elegant theory by Scottish physicist James Clark Maxwell in 1873. Maxwell's four equations describing the electromagnetic field are recognized as one of the great achievements of 19th century physics. Maxwell was able to calculate the speed of propagation of the electromagnetic field from his equations, and found it to be approximately equal to the speed of light. He then proposed that light is an electromagnetic phenomenon. Because electromagnetic fields can oscillate at any frequency, he concluded that visible light occupied only a very small portion of the frequency spectrum of electromagnetic radiation. The entire spectrum includes radio waves of low-frequency, high-frequency, very-high frequency, ultra-high frequency, and microwaves. At still higher frequencies are infrared radiation, visible light, ultraviolet radiation, x-rays, and gamma rays. All of these are fundamentally the same kind of waves, the only difference between them being the frequency of the radiation.

Now we ask, what is the electromagnetic field, anyway? Is it a physical object? To answer that question, we must understand what we mean by the term physical object. One definition is that it is anything that carries force, energy, and momentum. By this definition the electromagnetic field is a physical object because it carries force, energy, and momentum. However, this merely defines the electromagnetic field in terms of other things that require their own definitions. Force, energy, and momentum can only be defined in terms of the operations necessary to measure them and these operations require physical objects on which to make the measurements. Thus, all physical objects are defined in terms of other physical objects, so the definition is circular. This is another indication that the concept of objective reality is nothing but a concept.

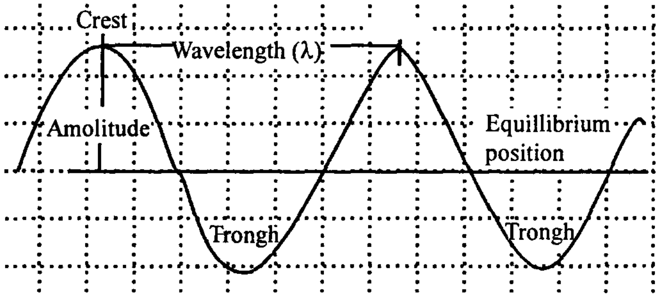


Fig. 1 Waves

These parameters are related by the following equation: v=λf

The electromagnetic spectrum contains electromagnetic waves of all frequencies and wavelengths:

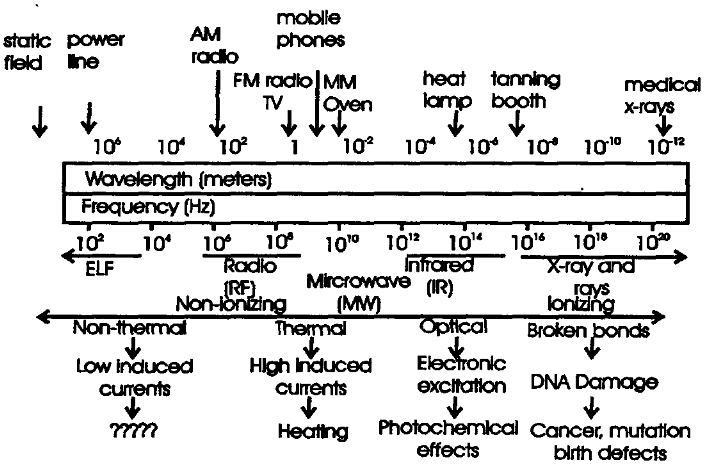


Fig.2 Electromagnetic waves of all frequencies and wavelengths

Waves

In the 1800s, it was known that light had a wave-like nature, and classical physics assumed that it was indeed a wave. Waves are traveling oscillations. Examples are water waves, which are traveling surface oscillations of water; and waves on a tightly stretched rope, which are traveling oscillations of the rope. Waves are characterized by three parameters: wavelength (l), oscillation frequency (f), and velocity (v).

It was not known what the oscillating medium was in the case of light, but it was given the name 'ether.' Maxwell had assumed that the ether provided an absolute reference frame with respect to which the velocity of any object or wave could be measured. In 1881, German-American physicist Albert Michelson and American physicist Edward Morley performed ground-breaking experiments on the velocity of light. They found that the velocity of light on the earth always had the same constant value regardless of the direction of motion of the earth about the sun. This violated the concept, which was prevalent at the time, that the measured velocity of any object, be it particle or wave, depends on the observer's velocity relative to the velocity of the other object. This concept is demonstrated in everyday life when our observation of another car's velocity depends on the velocity of our own car. Thus, the measured velocity of light relative to the ether was expected to depend on the direction of motion of the earth relative to the velocity of the ether. But, the constancy of the velocity of light meant that the concept of the ether had to be abandoned because the ether velocity could not be expected to change with the observer's velocity in just such a way that the velocity of light always had the same value. Thus, in the case of light waves, physicists concluded that there is no material medium that oscillates.

Relativity

Implicit in the preceding discussion of classical physics was the assumption that space and time were the contexts in which all physical phenomena took place. They were absolute in the sense that no physical phenomena or observations could affect them, therefore they were always fixed and constant.

In 1905, the German-Swiss-American physicist Albert Einstein revolutionized these ideas of time and space by publishing his theory of special relativity. In this theory, he abandoned the concept of the ether, and with that the concept of the absolute motion of an object, realizing that only relative motion between objects could be measured. Using only the assumption of the constancy of the velocity of light in free space, he showed that neither length nor time is absolute. This means that both length and time measurements depend on the relative velocities of the observer and the observed.

An observer standing on the ground measuring the length of an airplane that is flying by will obtain a minutely smaller value than that obtained by an observer in the airplane. An observer on earth comparing a clock on a spaceship with his clock on earth will see that the spaceship clock moves slower than the earth clock.

For an object having a mass, the special theory produced the famous relationship between the total energy (E) of the object, which includes its kinetic energy, and its mass (m):

$$E = mc^2$$

where c is the velocity of light in a vacuum. Einstein's special theory has been confirmed by thousands of experiments, both direct and indirect.

In Einstein's special theory of relativity, even though space and time were no longer separately absolute, they were still Euclidean. This mea1t that two straight lines in space-time which were parallel at one point always remained parallel no matter what the gravitational forces were.

In 1915, Einstein completed his greatest work, the general theory of relativity. Whereas the special theory deals with objects in uniform relative motion, i.e., moving with constant speed along straight lines relative to each other, the general theory deals with objects that are accelerating with respect to each other, i.e., moving with changing speeds or on curved trajectories. Examples of accelerating objects are an airplane taking off or landing, a car increasing or decreasing its speed, an elevator starting up or coming to a stop, a car going around a curve at constant speed, and the earth revolving around the sun or the moon revolving around the earth at constant speed.

A particularly important example of acceleration is that of an object free-falling in the earth's gravity. A free-falling object is one that is acted upon only by the gravitational force, without air friction or other forces. All free-falling objects at the same spot in the earth's gravitational field fall with the same acceleration, independent of the mass or material of the object. A free-falling object, such as an astronaut in a spaceship, does not experience a gravitational force (i.e., he/she experiences weightlessness), hence we can say that the acceleration of free-fall cancels out the gravitational force. Another way to state this fact is that a gravitational force is equivalent to an acceleration in the same direction. This is Einstein's famed equivalence postulate, which he used in discovering general relativity.

The equivalence postulate applies to all objects, even light beams. Consequently, the path of a light beam is affected by a gravitational field just like the trajectory of a baseball. However, because of the very high speed of the photons in a light beam (3×10^8 metres/second, or 186,000 miles/second), their trajectories are bent by only very tiny amounts in the gravitational fields of ordinary objects like the sun.

Because all types of objects are affected in exactly the same way by gravity, an equivalent way of looking at the problem is to replace all gravitational forces by curved trajectories. The curved trajectories are then equivalent to curving space itself! This is the second key concept that Einstein used in the general theory of relativity. The result is that the general theory replaces the concept of gravity with the curvature of space. The curvature of a light beam around an individual star or galaxy is very small and difficult to measure. Even the whole universe curves the trajectory of a light beam only a little.

Clear evidence that the force of gravity is nothing but a concept is given by the fact that it can be replaced by another concept, the concept of the curvature of space. Less clear is that the body sensations that we normally associate with the force of gravity are also nothing but concepts.

Speaking of the universe as a whole, what are the effects of curved space? The principal effect is that light beams no longer travel in straight lines. Hence, if two light beams start out parallel, they will eventually either converge or diverge. If they diverge, we say that space has negative curvature, and if they converge, we say that it has positive curvature. Zero curvature corresponds to parallel light beams always remaining parallel. This implies a Euclidean, or flat, space.

The electromagnetic field is nothing but a concept, we can now say that space is also nothing but a concept! It is a concept that allows us to conceptualize the separation of objects (which are nothing but concepts) and it allows us to predict the trajectories of light beams.

The curvature of the universe as a whole depends on the average mass density and on the expansion rate of the universe. The fact that the universe is expanding was discovered by American astronomer Edwin Hubble in 1929, 14 years after Einstein published his general theory of relativity.

Whether the space of our universe has positive or negative curvature is a matter for experimental determination. In practice, it is too difficult to do this by measuring the curvature of light beam trajectories, but the curvature can be calculated if the average mass density and the expansion velocity are known. The average mass density cannot easily be measured directly because we are unable to see matter that is not emitting light, so the average mass density in a galaxy, for example, must be calculated from the trajectories of the motion of visible stars in the galaxy. Such measurements indicate that there is a large amount of matter in the universe that does not shine with its own or reflected light. This is called dark matter.

Until 1998, it was thought that the universe was expanding at a constant rate, but in 1998 it was discovered that it is actually expanding at an accelerating rate rather than a constant one. This acceleration cannot be explained if the universe contains only ordinary and dark matter because these produce a gravitational force which is attractive, whereas an accelerating expansion requires a repulsive force. This repulsive force represents a 'dark energy' density in addition to the energy densities of ordinary and dark matter. Both dark matter and dark energy are presently being intensively investigated both theoretically and experimentally because they could be the result of new physical laws operating.

There are powerful theoretical reasons for believing that the curvature of our space is neither positive nor negative but is exactly zero. Zero curvature requires a certain value of the average mass density including both visible and dark matter. A larger value implies a positive curvature, and a smaller value implies a negative curvature. The density of visible matter by itself is not high enough to produce a zero or positive curvature.

In discovering the special theory of relativity, Einstein was heavily influenced by the positivism of Austrian natural philosopher Ernst Mach. Positivism is the philosophy that states that the only useful concepts are those that depend directly on empirical observation. This attitude is derived from the belief that the only objective, external reality that exists is one that can be directly observed, such as macroscopic objects. In inventing and explaining the special theory, Einstein followed the positivist approach and made extensive use of the empirical definitions of measurements of time and space, and he incorporated those definitions into the mathematics, which describe how length and time vary with the relative velocity of observer and observed. In this way, Einstein was able to avoid the concept of space except as being the context of measurements of length and time.

However, Einstein abandoned positivism when he developed the general theory of relativity, and it is unlikely that he could have developed it without doing so. His concept of general relativity depended essentially on an intuitive leap from the empirical operations of measuring the force of gravity and the accelerations of objects to a theoretical model of space which was curved and in which there were no gravitational forces. He likely could not have done this without believing that space was objectively real rather than being merely the context for making measurements of length and time.

In addition to curved space, a physicist who adhered to the positivist philosophy would not have discovered the electron, the atom, or quantum waves. Einstein's intuitive leap is an example of an essential aspect of the work of scientists. The individual experiments that scientists perform are always very specific to a particular problem in particular circumstances. Any attempt to comprehend the results of many such experiments on many similar topics would be futile without some kind of unifying model that is presumed to represent some aspect of the external, objective reality affecting those experiments.

For example, force fields are theoretical models of gravitational or electromagnetic forces, and curved space-time is a model of space-time that accounts for the gravitational force. There are other models that account for the weak and strong forces that act on elementary particles. And there are models of the nucleus, the atom, molecules, solids, crystals, and gases. All of these models are highly mathematical, because mathematics is the universal language of physics.

When a model is found that accurately accounts for experimental observations, there is a strong tendency to think of the model itself as the external, objective reality. Thus, both physicists and the general public routinely speak of elementary particles, nuclei, and atoms as being real objects, rather than simply as mathematical models. We shall see later that this tendency creates innumerable problems in trying to understand the true nature of Reality.

In classical physics, objects interact with each other through their force fields, which are also objects in external, objective reality. For example, the atoms and molecules in a solid, liquid, or gas are held together by the electromagnetic force. Charged particles also interact through the electromagnetic force. It turns out that all physical objects, which are nothing but concepts, interact with each other through their force fields, which are also nothing but concepts.

As revolutionary as Einstein's general theory of relativity was, it did nothing to change the belief that we as observers still live within the context of space-time even though space-time is no longer thought to be absolute and unchanging. This means, for example, that we as objects are still subject to the experience of separation and isolation from other objects, and to the experience of aging and the ultimate death of the body. It took an even more revolutionary theory, the quantum theory, to begin to shake these imprisoning beliefs.

2D Electron Gas

As was mentioned, quantum well structures have found important applications in novel semiconductor devices. In such structures, a thin region of a narrow gap semiconductor is sandwiched between layers of a wide band gap semiconductor or surrounded by a wide band gap semiconductor.

Let us first consider electrons in a narrow gap semiconductor layer, such as is shown in figure. If this layer is thin enough, the motion of carriers in the direction perpendicular to the heterointerfaces is quantized, meaning that this motion involves discrete (quantum) energy levels. In this case, electrons propagating in the narrow gap semiconductor are often referred to as a two-dimensional electron gas. Electrons in an unrestricted semiconductor are sometimes called a three-dimensional electron gas. Electrons propagating often called a quantum wire are called a one-dimensional electron gas.

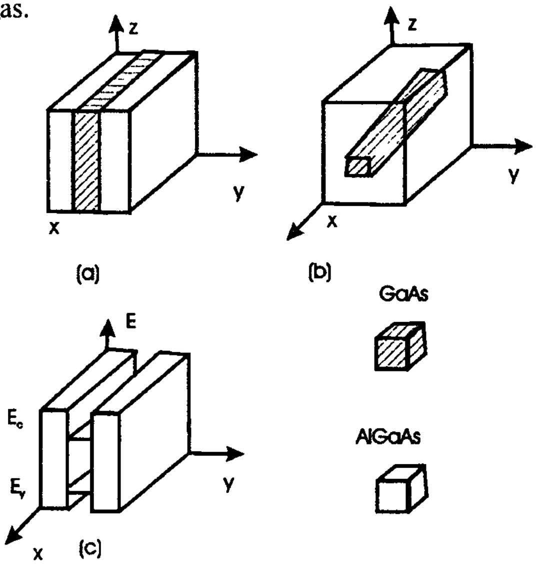


Fig. 3 (a) GaAs layer sandwiched between AlGaAs layers, (b) GaAs region surrounded by AlGaAs, and (c) corresponding band diagram. (AlGaAs has a wider energy gap than GaAs).

The lowest energy levels for a square potential well can be estimated as follows:

$$E_j - E_c = \frac{\pi^2 \hbar^2}{2m_n d^2} j^2$$

Here j is the quantum number labelling the levels, and d is the thickness of the quantum well. For the quantization to be important, the difference between the levels should be much larger then the thermal energy $k_B T$, that is,

$$\frac{\pi^2 \hbar^2}{2m_n d^2} >> K_B T$$

Using this condition, we find, for example, that in GaAs where $m_n/m_e = 0.067$, the levels are quantized at room temperature when d =150 E.

In the direction parallel to the heterointerfaces, the electronic motion is not restricted. Hence, the wave function for a two-dimensional electron gas can be presented as

$$\psi = f(y)\exp(ik_x x + ik_z z)$$

where f (y) may be approximated by:

$$f(y) = \left(\frac{2}{d}\right)^{1/2} \sin\left(\frac{\pi j}{d} y\right)$$

The term $\exp(ik_x x + ik_z z)$ in the wave function describing the electronic motion in directions x and z is similar to that of free electrons. This is understandable since electrons move freely in these directions. The dependence of the electron energy on the wave vector for a two-dimensional electron gas is given by

$$E - E_j = \frac{\hbar\left(k_x^2 + k_z^2\right)}{2m_n}$$

The k_y-component is absent in last equation since the motion in the y-direction is quantized. Each quantum level, E_j, corresponds to an energy subband.

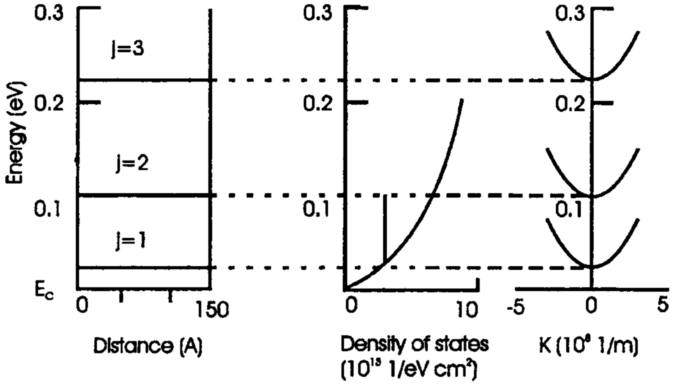


Fig.4 Energy Levels (bottoms of subbands), Density of States for Quantum Well Structure, and Energy Versus $k = (k_x^2 + k_z^2)^{1/2}$ for Two-dimensional Electron Gas in GaAs Quantum Well.

The density of states for each subband can be found using an approach similar to that used above for a three-dimensional density of states, that is, by counting the number of states with wave vectors k between k and dk. The corresponding area in k-space is equal to 2πkdk. The density of allowed states is equal to the number of allowed values of k in this area in k-space times two (the factor of 2 takes into account the two possible values of spin). The density of allowed points in k-space for the unit size sample is $1/(2\pi)^2$. Hence, the total number of states with values of k between k and k + dk is

$$dN = \frac{2 \times 2\pi k dk}{(2\pi)^2}$$

Taking into account that

$$k = \frac{\sqrt{2m_n(E - Ej)}}{\hbar}$$

where E_j is the bottom of the j^{th} subband and

$$kkdk = m_n dE / \hbar^2$$

we obtain

$$dN = \frac{4\pi\left[2m_n\left(E - E_j\right)h\right]^{1/2} 2m_n dE / h^2}{2\left[2m_n(E - E)h^2\right]^{1/2}(2\pi)^2} = DdE$$

where the density of states, D, for one subband is given by $D = \frac{m_n}{\pi h^2}$

The states of the first (bottom) subband overlap with the states of the second (from the bottom) subband for energies larger than the second energy level, and so on. As a consequence, the overall density of states has a 'staircase' shape. With an increase in the well thickness, d, the steps in Fig. 4 gradually decrease and merge into an envelope parabolic function, which is equal to the three-dimensional density of states function multiplied by d.

1D Electron Gas

Let us now consider a one-dimensional quantum wire where electron motion in two directions (y and z) is quantized and in one direction (x) electrons are free to move. The wave function y(x, y, z) and dispersion relation $E_{n1,n2.k}$ are now given by

$$\psi = f(y)f(z)\exp(ik_x - x)$$

where f(y) and f(z) are functions localized within the cross section of the quantum wire,

$$E - E_c = E_i 1_i 2 + \frac{\hbar^2 k^2 x}{2m_n} = \frac{\pi^2\hbar^2}{2m_n d_y^2} i_1^2 + \frac{\pi^2\hbar^2}{2m_n d_z^2} i_2^2 + \frac{\hbar^2 k_x^2}{2m_n}$$

Where i_1 and i_2 are quantum numbers related to quantization in the y and z directions, dy and dz are dimensions of the quantum wire in the y and z directions, and where we assume that the dispersion relation for the electron energy in each subband is parabolic.

The density of states for a one-dimensional subband can be found by using an approach similar to that used above for a two-dimensional density of states, that is, by counting the number of states with wave vectors k between k and dk. For a one-dimensional system, the density of allowed points in k-space for a unit size sample is $1/(2\pi)$. Hence, the total number of states with k between k and k + dk is

$$dN = 2 \times 2\frac{dk}{2\pi} = \frac{2dk}{\pi}$$

[An additional factor of 2 appears here because there are two directions of k: positive and negative.] Using equation,

$$dk = \frac{m_n^{1/2} dE}{\hbar\sqrt{2\left(E - E_{i1,i2}\right)}}$$

where $E_{i1,i2}$ is to the bottom of the subband corresponding to the quantum numbers i_1 and i_2. Hence,

$$dN = \frac{\sqrt{2} m_n^{1/2} dE}{\pi\hbar\sqrt{E - E_{i1,i2}}} = \Omega dE$$

Where the density of states is given by

$$\Omega(E) = \frac{\sqrt{2} m_n^{1/2}}{\pi\hbar\sqrt{E - E_{i1,i2}}}$$

Below we compare the densities of states for three-dimensional, two-dimensional and one-dimensional electron gases in GaAs. Consider only the two lowest subbands for the two-dimensional and one-dimensional electron gases.

If we choose the cross section of the GaAs quantum wire containing the one-dimensional gas to be equal to 100 E ×100 E. then the lowest energies in the two lowest subbands are equal to 0.112 eV and 0.280 eV. The dependencies of these densities of states on energy.

Of course, the densities of states for one-, two-, and three-dimensional electron gases have different dimensions. If we make, for example, 105 parallel identical quantum wires per cm then the two-dimensional density of states in all these wires will be 105 greater than W and will be more comparable to D. In similar way, we can consider many identical parallel layers containing two-dimensional gas and then multiplying D by the number of layers per cm, we can obtain the three-dimensional density of states, which will be more comparable to g.

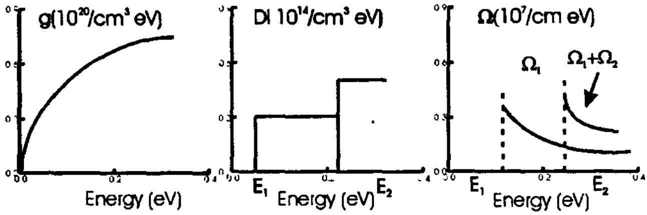


Fig. 5 Densities of States Versus Energy for Three-dimensional (g), Two-dimensional (D), and One-dimensional (W) Electron Gases in GaAs Conduction band. Only the Two lowest Subbands are accounted for Two-dimensional and One-dimensional Electron Gases.

Once we find the densities of states we can calculate the electron concentration in the conduction band. However, energy states in the valence band may play an equally important role.

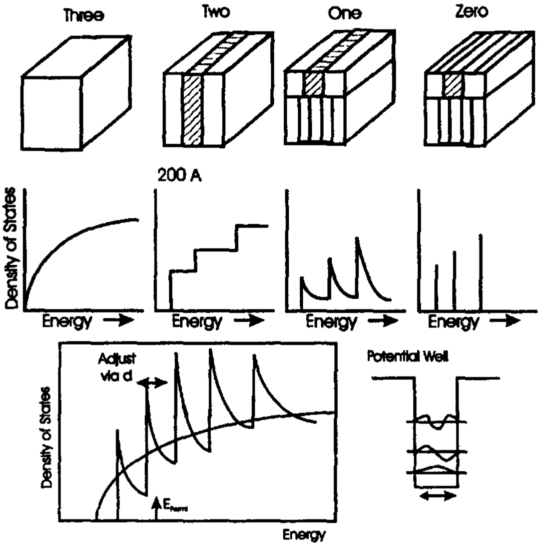


Fig. 6 Energy states in the valence band

Tailoring Electronic Properties of Materials by Nanostructuring

Electrons confined to nanostructures give rise to low-dimensional quantum well states, which modify the density of states. States at the Fermi level trigger electronic phase transitions, such as magnetism and superconductivity.

How fine does a solid need to be structured to have an impact on its electronic properties? The wavefunction of electrons is going to change when they are confined to dimensions comparable with their wavelength. As an estimate we may see the Fermi wavelength of a simple free-electron gas. It decreases with increasing carrier density. Therefore, confinement and quantization phenomena are visible in semiconductors already at dimensions greater than 200 nm, whereas in metals they typically are seen at 1 nm.

In fact, the Fermi wavelength of typical metals has atomic dimensions, but beat frequencies with the lattice can be an order of magnitude larger. A related way of reasoning considers the formation of low-dimensional electronic states by quantization.

Confining electrons to small structures causes the continuous bulk bands to split up into discrete levels, for example quantum well states in a slab. For N atomic layers in the slab there are N levels. In order to exhibit two-dimensional behaviour there should be only a single level within ±kT of the Fermi level. Several levels within the Fermi cut-off would already approach a three-dimensional continuum. For a coarse estimate of the corresponding slab thickness, one may set the energy E of the lowest level equal to kT. For room temperature ($E = kT = 0.026$ eV), one obtains a de Broglie wavelength $l = h/p = h/(2mE)^{1/2} = 1.23 \text{ nm} /(E/eV)^{1/2} = 8$ nm, which is comparable with the spatial extent of the lowest quantum state. Thus, both the high electron density and the requirement of room-temperature operation for quantum devices point to dimensions of a few nanometres.

Superlattice Devices

To use a semiconductor superlattice based on a periodic structure of alternating layers of semiconductor materials with wide and narrow band gaps. The first superlattices were fabricated using an AlGaAs/GaAs material system.

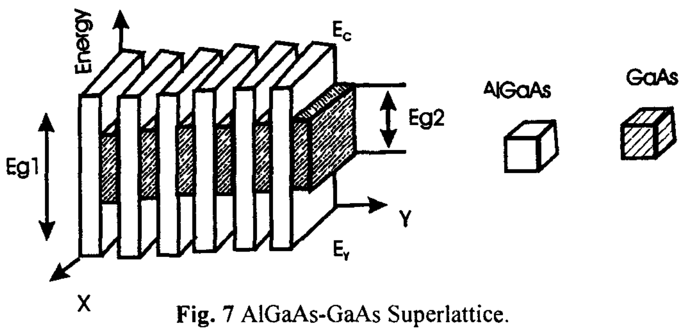


Fig. 7 AlGaAs-GaAs Superlattice.

Let us first review what is happening in one individual quantum well formed in the GaAs layer sandwiched between two AlGaAs barrier layers. If the GaAs layer thickness is small enough, then electronic motion in the quantum well is quantized in the direction perpendicular to the heterointerfaces. The carriers move freely in the direction parallel to the heterointerfaces so that the wave function is proportional to exp[i(k_xx + k_zz)f(y). Here k_x and k_z are components of the wave vector in the plane of the superlattice, x and z are the coordinates in the superlattice plane, and y is the coordinate perpendicular to the superlattice plane. As was discussed, each energy level found for a quantum well from the solution of the one-dimensional Schrodinger equation corresponds to a subband of states with the density of states, D, in each subband given by

$$D = \frac{m_n}{\pi h^2}$$

If the thickness of the wide-band-gap barriers layers is small enough so that electrons may tunnel through, then the situation becomes similar to what happens when individual atoms are brought together in a crystal. In this case, individual levels in the quantum wells are split into bands (called the minibands). In a crystal, the periodic atomic potential leads to band formation. In a superlattice, an artificial, human-made periodical potential causes the formation of minibands.

To create an artificial periodic potential in a semiconductor crystal using periodic n-type and p-type doped layers. Such a superlattice is called a doping superlattice.

Superlattice structures have been be used in field effect transistors where several quantum wells provide parallel conducting channels, increasing the device current carrying capabilities and, hence, the output power. Superlattices are also used for photodetectors and for novel light-emitting diodes.

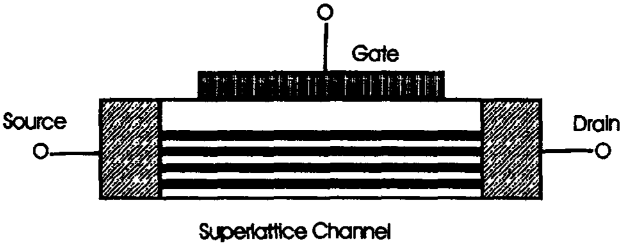


Fig. 8 Heterostructure Field Effect Transistor with a Superlattice Channel.

In heterostructure devices, superlattice buffers are used to create an intermediate layer between a substrate and an active layer. This allows us to alleviate strain caused by lattice constant mismatch and to obtain a much better quality active layer material.

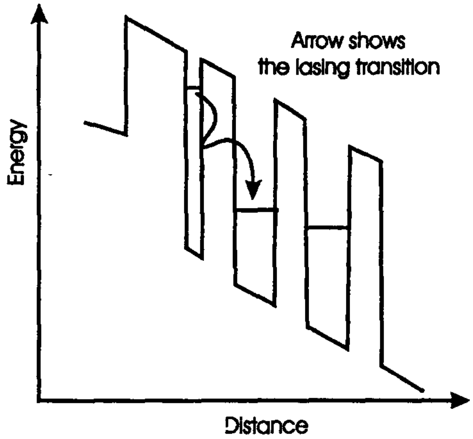


Fig. 9 Active Region of Quantum Cascade Laser.

2

Max Planck's Revolutionary Hypothesis

Quantum Revolutioı.

Over three hundred years ago, Sir Isaac Newton revolutionized the study of the natural world by putting forth laws of nature that were stated in mathematical form for the first time.

By the start of the 20th century, physicists had worked with Newton's laws so thoroughly that some of them thought that they were coming to the end of physics. In their opinion, not much was left to do to make physics a complete system. Little did they know that the world they described was soon to be understood in a completely different way. The *quantum revolution* was about to happen.

This revolution was begun by a very unlikely person, a physicist named Max Planck, who was very conservative in all his views. It speaks well of Planck's intellectual honesty that he was able to accept the reality of what he discovered, even though he found the consequences of his discoveries distasteful and unpleasant for the rest of his life.

Born in 1853, Max Planck came from a conservative and respectable family in Kiel, Germany. Young Max was very bright, and had a variety of fields from which to choose to study for his professional life. Planck chose physics because he felt that it was the field in which he was most likely to do some original work. At the young age of 21, he received his doctorate in physics from the University of Munich.

Planck was investigating the properties of heat- and light-emitting bodies. Classical physics had theories which predicted that the brightness of a body *increases continuously* as the frequency of its electromagnetic radiation is increased.

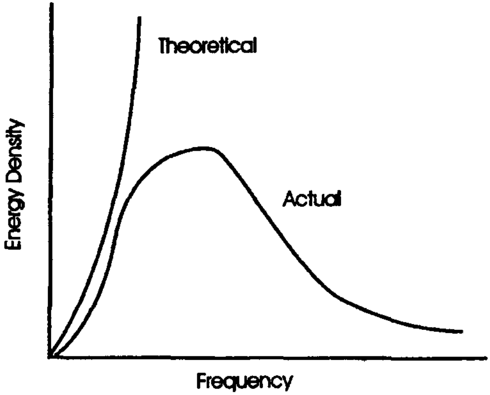


Fig. 1 Intensity vs. Frequency Plot.

Unfortunately, experiments revealed a totally different picture. *The brightness did increase initially, but only up to a limit. Then, actually, it began to fall.* We thus get a bell-shaped curve if we plot frequency against brightness.

Besides, another observation was made: *as bodies become hotter, their maximum brightness shifts towards higher frequencies.* This is why an object, heated to 300-400°C, emits mostly infrared or heat waves. As the temperature is increased, the object appears to be red, then orange, and finally white or even blue.

Classical theories totally failed to explain this discrepancy between the known facts and the observations. Then, in the winter of 1900, Max Planck found a solution to this problem. Planck ushered in the quantum era by making a bizarre assumption:

Emission and absorption of energy can occur only in discrete amounts.

This might seem totally unsurprising to you, but believe me, it shook the scientists of that period. Planck himself did not know he would end up with this statement!

Imagine for a moment that you are a sculptor, and you have obtained a piece of stone in the shape of a cube. To begin your sculpture, you take a chisel and place its edge against the stone, and then strike the chisel with a hammer. What do you imagine would happen? I think you would imagine that a piece of the stone would be split off, as well as some smaller splinters and pieces of stone. Imagine instead that when you struck the stone, it broke into hundreds of small cubes, each one of them exactly the same size, 3 centimetres per side. Wouldn't you be surprised, even shocked? Imagine furthermore that no matter how hard you tried, these smaller cubes could not be broken into smaller pieces at all!

We think that the reaction that a sculptor would have in such a circumstance would be similar to what Planck and other physicists felt upon discovering that energy only occurred in discrete amounts. It was a completely unexpected discovery, and yet it was only the beginning of what would come later.

Planck called these discrete lumps as *quanta*. This was against the entire worldview that had been built from the time of Newton onward. In the physics that had been built up since the time of Newton, and indeed in the minds of most thinkers before Newton, matter and energy were thought to be smooth and continuous. Even by the time of Planck, the idea that matter could be ultimately broken down into tiny indivisible 'atoms' was only held by a few physicists.

The Beginning of Quantum Physics

Physicists measure the spectrum (the intensity of light as a function of wavelength, or colour) of a light source in a spectrometer. The figure below shows a schematic drawing of a simple prism spectrometer. White light comes in from the left and the prism disperses the light into its colour spectrum.

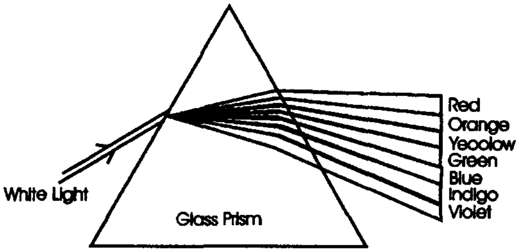


Fig. 2 Spectrum

In the late 1800s, physicists were making accurate measurements of the spectra of the emissions from black bodies (objects which are opaque, or highly absorbing, to the light they emit). Good examples of black bodies are the sun, the filament of an incandescent lamp, and the burner of an electric stove. The colour of a black body depends on its temperature, a cool body emitting radiation of long wavelengths, i.e., in the radio frequency range or in the infrared which are invisible to the eye, a warmer body emitting radiation which includes shorter wavelengths and appearing deep red, a still warmer body emitting radiation which includes still shorter wavelengths and appearing yellow, and a hot body emitting even shorter wavelengths and appearing white. The emissions are always over a broad range of colours, or wavelengths, and their appearance is the net result of seeing all of the colours at once.

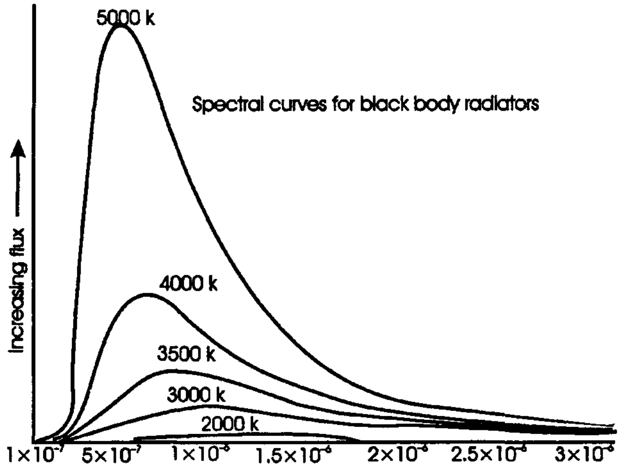


Fig. 3 Spectral Curves for Black Body Radlators

Classical physics could not explain the spectra of black bodies. It predicted that the intensity of emitted light should increase rapidly with decreasing wavelength without limit (the 'ultraviolet catastrophe'). In the fig.4, the curve labelled 'Rayleigh-Jeans law' shows the classically expected behaviour.

However, the measured spectra actually showed an intensity maximum at a particular wavelength, while the intensity decreased at wavelengths both above and below the maximum. In order to explain the spectra, in 1900 the German physicist Max Planck was forced to make a desperate assumption for which he had no physical explanation. As with classical physics, he assumed the body consisted of vibrating oscillators (which were actually collections of atoms or molecules).

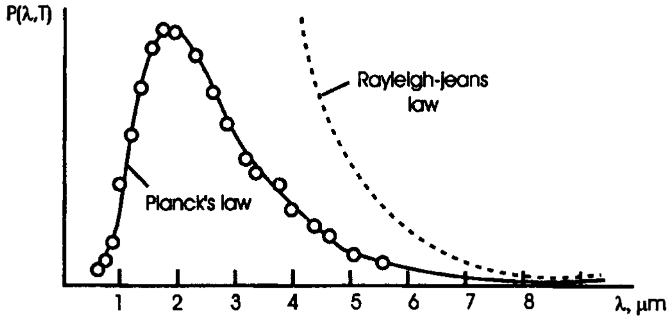


Fig. 4 Rayleigh-Jeans law

However, in contrast to classical physics, which assumed that each oscillator could absorb an arbitrary amount of energy from the radiation or emit an arbitrary amount of energy to it, Planck was forced to assume that each oscillator could receive or emit only discrete, quantized energies (E), such that

$E = hf$ (Planck's formula)

where h (Planck's constant) is an exceedingly small number whose value we do not need to present here, and f is the frequency of vibration of the oscillator (the number of times it vibrates per second). Each oscillator is assumed to vibrate only at a fixed frequency (although different oscillators in general had different frequencies), so if it emitted some radiation, it would lose energy equal to hf, and if it absorbed some radiation, it would gain energy equal to hf. Planck did not understand how this could be, he merely made this empirical assumption in order to explain the spectra. The figure above shows Planck's prediction; this agreed with the measured spectra.

Also in the late 1800s, experimental physicists were measuring the emission of electrons from metallic objects when they shined light on the object. This is called the *photoelectric effect.* These experiments also could not be explained using classical concepts. These physicists observed that emission of electrons occurred only for light wavelengths shorter than a certain threshold value that depended on the metal. Classically, however, one expected that the emission should not depend on wavelength at all, but only on intensity, with greater intensities yielding more copious emission of electrons.

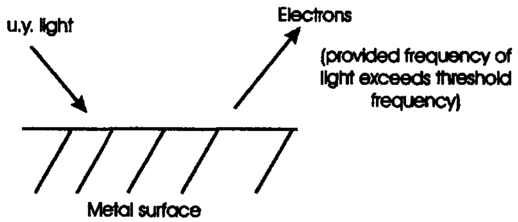


Fig. 5 Emission of electron

In one of a famous series of papers in 1905, Einstein explained the photoelectric effect by starting with Planck's concept of quantized energy exchanges with light radiation, and making the startling assumption that these quantized exchanges were a direct result of the quantization of light itself, i.e., light consisted of discrete bundles of energy called photons, rather than the continuous waves which had always been assumed by classical physicists. However, these bundles still had a wave nature, and could be characterized by a wavelength, which determined their colour. He also used Planck's relationship between energy and frequency (E = hf) to identify the energy of the photon, and he used the relationship between velocity, frequency, and wavelength that classical physics had always used (v = lf, where now v = c = velocity of light).

In classical physics, the electromagnetic field connects charged particles to each other. In quantum physics, the force fields of classical physics are quantized, and the quanta of the fields then become the force carriers. For example, photons are the quanta of the electromagnetic field. In quantum physics, it is the photons that connect charged particles to each other.

The Development of Quantum Mechanics

In addition to measuring the spectra of blackbody radiation in the 19th century, experimental physicists also were familiar with the spectra emitted by gases through which an electrical discharge (an electric current with enough energy to strip some of the electrons from the atoms of the gas) was passing. Examples of such discharges are the familiar neon sign, in which the gas is neon; and the fluorescent light bulb, in which the gas is mercury vapour (the fluorescent light bulb has special coatings on the inner walls which change the spectrum of the light). The spectra of such light sources consist of emissions at discrete, separated wavelengths, rather than over a continuous band of wavelengths as in blackbody spectra. These spectra are called line spectra because of their appearance when they are viewed with a spectrometer.

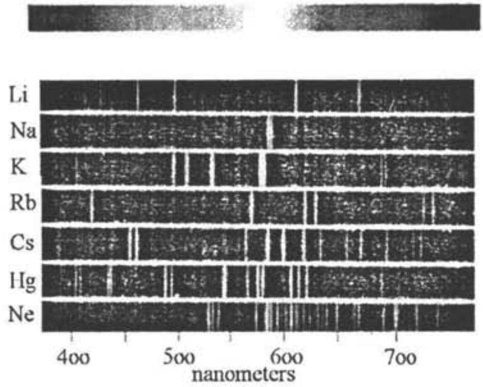


Fig. 6 Black Body and Line Spectra Black Body

When classical physics was applied to such a model of the atom, it predicted that the electrons could not remain in stable orbits about the nucleus, but would radiate away all of their energy and fall into the nucleus, much as an earth satellite falls into the earth when it loses its kinetic energy due to atmospheric friction. In 1913, after Danish physicist Niels Bohr had learned of these results, he constructed a model of the atom that made use of the quantum ideas of Planck and Einstein. He proposed that the electrons occupied discrete stable orbits without radiating their energy. The discreteness was a result of the quantization of the orbits, with each orbit corresponding to a specific quantized energy for the electron. The electron was required to have a certain minimum quantum of energy corresponding to a smallest orbit; thus, the quantum rules did not permit the electron to fall into the nucleus. However, an electron could jump from a higher orbit to a lower orbit and emit a photon in the process. The energy of the photon could take on only the value corresponding to the difference between the energy of the electron in the higher and lower orbits. An electron could also absorb a photon and jump from a lower orbit to a higher orbit if the photon energy equaled the difference in orbit energies. Bohr applied his theory to the simplest atom, the hydrogen atom, which consists of one electron orbiting a nucleus of one proton.

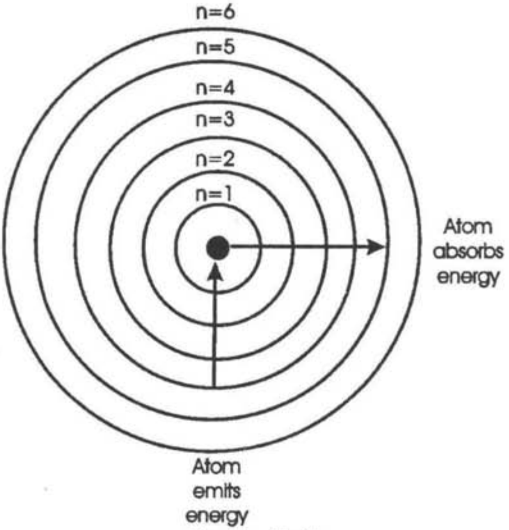


Fig. 7 n Shell

The theory explained many of the properties of the observed line spectrum of hydrogen, but could not explain the next more complicated atom, that of helium, which has two electrons. Nevertheless, the theory contained the basic idea of quantized orbits, which was retained in the more correct theories that came later.

In the earliest days of the development of quantum theory, physicists, such as Bohr, tried to create physical pictures of the atom in the same way they had always created physical pictures in classical physics. However, although Bohr developed his initial model of the hydrogen atom by using an easily visualized model, it had features that were not understood, and it could not explain the more complicated two-electron atom. The theoretical breakthroughs came when some German physicists who were highly sophisticated mathematically, Werner Heisenberg, Max Born , and Pascual Jordan largely abandoned physical pictures and created purely mathematical theories that explained the detailed features of the hydrogen spectrum in terms of the energy levels and the intensities of the radiative transitions from one level to another. The key feature of these theories was the use of matrices instead of ordinary numbers to describe physical quantities such as energy, position, and momentum.

The step of resorting to entirely mathematical theories that are not based on physical pictures was a radical departure in the early days of quantum theory, but today in developing the theories of elementary particles it is standard practice. Such theories have become so arcane that physical pictures have become difficult to create and to picture, and they are usually developed to fit the mathematics rather than fitting the mathematics to the physical picture. Thus, adopting a positivist philosophy would prevent progress in developing models of reality, and the models that are intuited are more mathematical than physical.

Nevertheless, in the early 1920s some physicists continued to think in terms of physical rather than mathematical models. In 1923, French physicist Louis de Broglie reasoned that if light could behave like particles, then particles such as electrons could behave like waves, and he deduced the formula for the wavelength of the waves:

$$\lambda = h/p$$

where p is the momentum (mass x velocity) of the electron. Experiments subsequently verified that electrons actually do behave like waves in experiments that are designed to reveal wave nature.

In physics, if there is a wave then there must be an equation that describes how the wave propagates in time. De Broglie did not find it, but in 1926 German physicist Erwin Schrodinger discovered the celebrated equation that bears his name. The Schrodinger equation allows us to calculate precisely the Schrodinger wave at all points in space at any future time if we know the wave at all points in space at some initial time. In this sense, even quantum theory is completely deterministic.

Schrodinger verified his equation by using it to calculate the line emission spectrum from hydrogen, which he could do without really understanding the significance of the waves. In fact, Schrodinger misinterpreted the waves and thought they represented the electrons themselves. However, such an interpretation could not explain why experiments always showed that the photons emitted by an atom were emitted at random rather than predictable times, even though the average rate of emission could be predicted from both Heisenberg's and Schrodinger's theories. It also could not explain why, when an electron is detected, it always has a well-defined position in space, rather than being spread out over space like a wave.

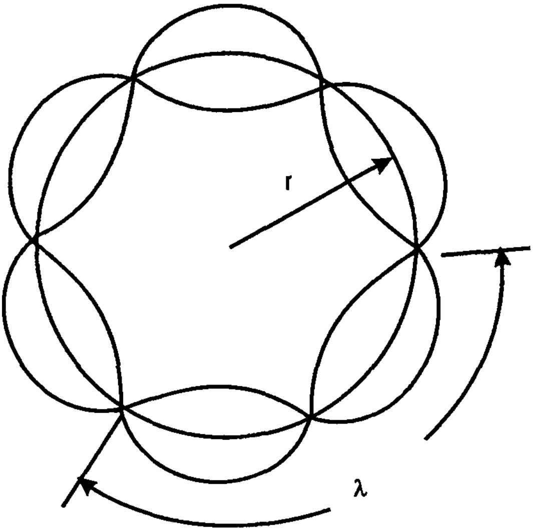


Fig. 8 Schrodinger misinterpreted the waves

The proper interpretation was discovered by German physicist Max Born, who suggested that the wave (actually, the absolute value squared of the amplitude or height of the wave, at each point in space) represents the probability that the electron will appear at that specified point in space if an experiment is done to measure the location of the electron. Thus, the Schrodinger wave is a probability wave, not a wave that carries force, energy, and momentum like the electromagnetic wave. Born's interpretation introduces two extremely important features of quantum mechanics:

1. From the wave we can calculate only probabilities, not certainties (the theory is probabilistic, not deterministic).
2. The wave only tells us the probability of finding something if we look, not what is there if we do not look (quantum theory is not a theory of objectively real matter although Born thought the Schrodinger wave was objectively real).

The first feature violates the second fundamental assumption of classical physics, i.e., that both the position and velocity of an object can be measured with no limits on their precision except for those of the measuring instruments. The second feature violates the first fundamental assumption of classical physics, i.e., that the objective world exists independently of any observations that are made on it.

Uncertainty and Complementarity

As Born proposed, quantum theory is intrinsically probabilistic in that in most cases it cannot predict the results of individual observations. However, it is deterministic in that it can exactly predict the probabilities that specific results will be obtained. Another way to say this is that it can predict exactly the average values of measured quantities, like position, velocity, energy, or number of electrons detected per unit time in a beam of electrons, when a large number of measurements are made on identical electron beams. It cannot predict the results of a single measurement. This randomness is not a fault of the theory—it is an intrinsic property of nature. Nature is not deterministic in the terms thought of in classical physics.

Another feature of the quantum world, the world of microscopic objects, is that it is intrinsically impossible to measure simultaneously both the position and momentum of a particle. This is the famous uncertainty principle of Heisenberg, who derived it using the multiplication rules for the matrices that he used for position and momentum. For example, an apparatus designed to measure the position of an electron with a certain accuracy in following diagram. The hole in the wall ensures that the positions of the electrons as they pass through the hole are within the hole, not outside of it.

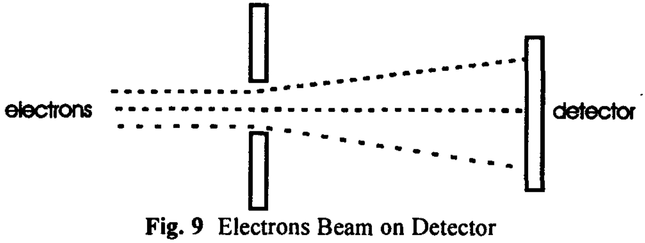


Fig. 9 Electrons Beam on Detector

So far, this is not different from classical physics. However, quantum theory says that if we know the position q of the electron to within an accuracy of Δq (the diameter of the hole), then our knowledge of the momentum p (= mass x velocity) at that point is limited to an accuracy Δp such that

(Δp)(Δq)>h (Heisenberg uncertainty relation)

In other words, the more accurately we know the position of the electron (the smaller Δq is), the less accurately we know the momentum (the larger Δp is). Since momentum is mass times velocity, the uncertainty in momentum is equivalent to an uncertainty in velocity. The uncertainty in velocity is in the same direction as the uncertainty in position. In the drawing above, the uncertainty in position is a vertical uncertainty. This means that the uncertainty in velocity is also a vertical uncertainty. This is represented by the lines diverging (by an uncertain amount) after the electrons emerge from the hole (uncertain vertical position) rather than remaining parallel as they are on the left.

Likewise, an experiment designed to measure momentum with a certain accuracy will not be able to locate the position of the particle with better accuracy than the uncertainty relationship allows.

Notice that in the uncertainty relationship, if the right side equals zero, then both Δp and Δq can also be zero. This is the assumption of classical physics, which says that if the particles follow parallel trajectories on the left, they will not be disturbed by the hole, and they will follow parallel trajectories on the right.

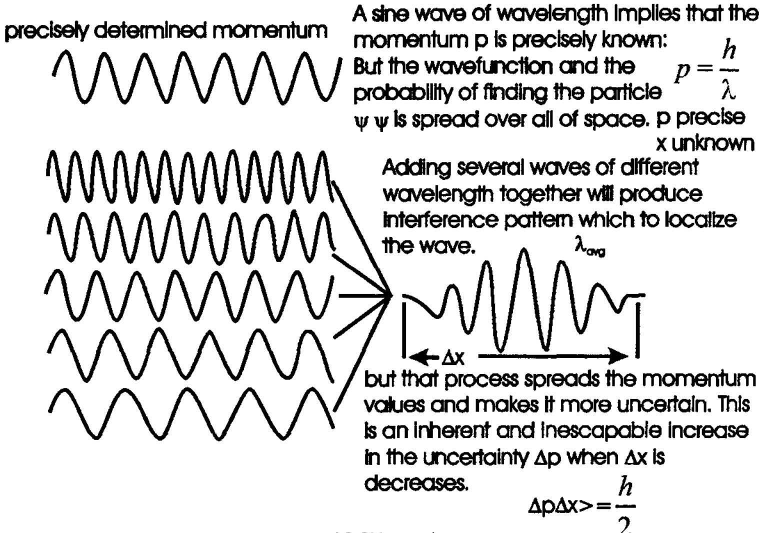
If we divide both sides of the uncertainty relation by the mass m of the particle, we obtain

$$(\Delta v)(\Delta q)>h/m$$

Here we see that the uncertainties in velocity v or position q are inversely proportional to the mass of the particle. Hence, one way to make the right side effectively zero is to make the mass very large. When numbers are put into this relationship, it turns out that the uncertainties are significant when the mass is microscopic, but for a macroscopic mass the uncertainty is unmeasurably small. Thus, classical physics, which always dealt with macroscopic objects, was close to being correct in assuming that the position and velocity of all objects could be determined arbitrarily accurately.

The uncertainty principle can be understood from a wave picture. A wave of precisely determined momentum corresponds to an infinitely long train of waves, all with the same wavelength, as is shown in the first of the two wave patterns below. This wave is spread over all space, so its location is indeterminate.

A wave of less precisely determined momentum can be obtained by superposing waves of slightly different wavelength (and therefore slightly different momentum) together, as is shown in the second of the two patterns above.



A sine wave of wavelength implies that the momentum p is precisely known: But the wavefunction and the probability of finding the particle ψ ψ is spread over all of space. p precise x unknown $p = \frac{h}{\lambda}$

Adding several waves of different wavelength together will produce interference pattern which to localize the wave.

but that process spreads the momentum values and makes it more uncertain. This is an inherent and inescapable increase in the uncertainty Δp when Δx is decreases. $\Delta p \Delta x >= \frac{h}{2}$

Fig. 10 Wavepicture

This results in a wave packet with a momentum spread Δp (uncertainty Δp), but which is bunched together into a region of width Δx (uncertainty Δx) instead of being spread over all space.

The uncertainty relation is closely related to the complementarity principle, which was first enunciated by Bohr. This principle states that quantum objects (objects represented by quantum wavefunctions) have both a particle and a wave nature, and an attempt to measure precisely a particle property will tend to leave the wave property undefined, while an attempt to measure precisely a wave property will tend to leave the particle property undefined. In other words, particle properties and wave properties are complementary properties. Examples of particle properties are momentum and position. Examples of wave properties are wavelength and frequency. A precise measurement of momentum or position leaves wavelength or frequency undefined, and a precise measurement of wavelength or frequency leaves momentum or position undefined.

Complementarity and uncertainty strongly imply that the electron (or any other 'particle') is neither a particle nor a wave. If so, what is it? So far, we have neglected the role of the observer in all measurements. When we take that into account, in fact there are actually neither particles nor waves. But if there are no observed objects, and there are only observations, then there is no external objective reality.

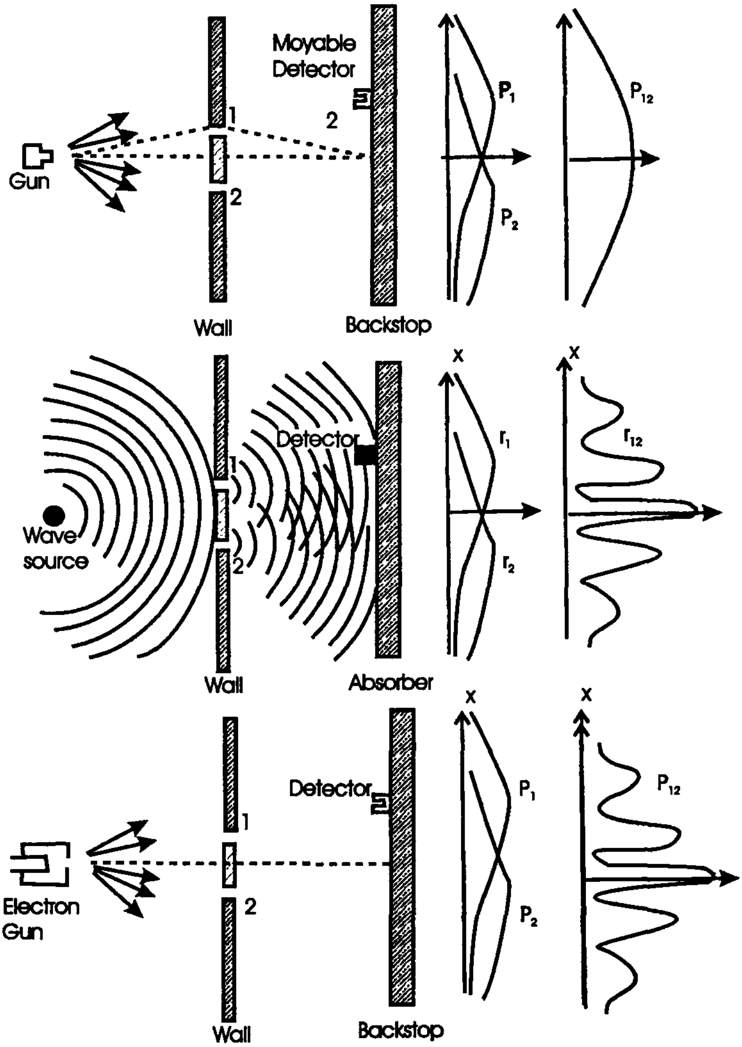


Fig. 11 Interference of Wave

Waves and Interference

Let us review the concept of the probability wave. The quantum wave does not carry energy, momentum, or force. Its sole interpretation is that from it we can calculate the probability that a measurement will yield a particular result, e.g., that photographic film will measure a specific position of an electron in an electron beam, or that a Geiger counter will yield a specific number of gamma rays from a radioactive source. It is only during a measurement that a particle appears. Prior to the measurement, what exists is not something that can be determined by either quantum theory or by experiment, so it is a metaphysical question, not a question of physics. However, that does not mean that the metaphysical answer does not have considerable impact in both the scientific world and one's personal world.

Suppose we do an experiment in which machine gun bullets are fired at a wall with two holes in it. The probability P_{12} of finding a bullet from either hole at the backstop to the right of the wall is equal to the probability P_1 of finding a bullet from hole

1. Plus the probability P_2 of finding a bullet from hole.
2. The probability distributions are simply additive.

When we are dealing with waves, we have a different rule. The superposition principle is one that is obeyed by all waves in material media provided their amplitudes are not too great, and is rigorously obeyed by all electromagnetic waves and quantum waves. It says that the net wave amplitude or height at any point in space is equal to the algebraic sum of the heights of all of the contributing waves. In the case of water waves, we can have separate waves due to the wake of a boat, the splashing of a swimmer, and the force of the wind. At any point on the surface of the water, the heights of the waves add, but it is important to include the sign of the height, which can be negative as well as positive.

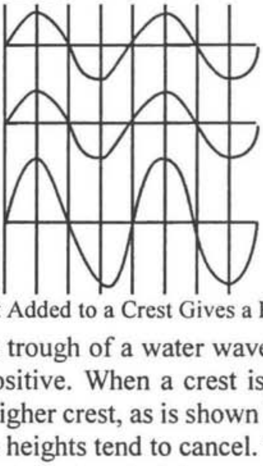


Fig. 12 Crest Added to a Crest Gives a Higher Crest.

The height of the trough of a water wave is negative while the height of a crest is positive. When a crest is added to a crest, the heights add to give a higher crest, as is shown below. When a trough is added to a crest, the heights tend to cancel. They cancel exactly if the heights of the crest and the trough are exactly equal but opposite in sign. When a trough is added to a trough, a deeper trough is created. When a crest is not lined up with either a crest or a trough, an intermediate wave is created.

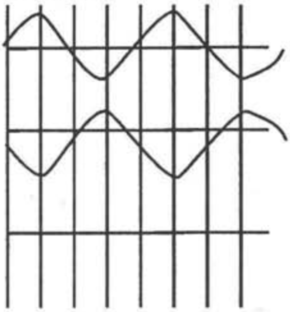


Fig. 13 Crest Added to a Trough Gives Cancellation.

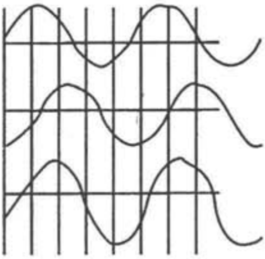


Fig. 14 Two Waves Added out of Phase Give an Intermediate Wave.

The superposition principle leads to the phenomenon of interference. The superposition, or sum, of two waves with the same wavelength at a point in space where both waves have either positive or negative heights results in a summed wave with positive or negative height greater than that of either one. This is called constructive interference. If the individual heights have opposite signs, the interference is destructive, and the height of the summed wave is smaller than the largest height of the two.

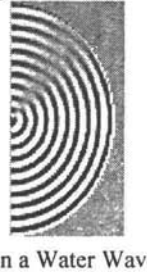


Fig. 15 Looking down on a Water Wave. The Bright Lines are Crests, the Dark ones are Troughs.

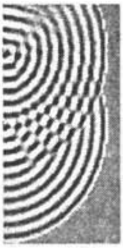


Fig. 16 Interference of Two Water Waves.

Crests added to crests form higher crests. Troughs added to troughs form deeper troughs.

An important measurable property of classical waves is power, or intensity (power per unit area). Power is proportional to the square of the wave amplitude, and is always positive. Interference of classical waves is illustrated in the middle panel of Fig., where the intensity I_{12} at the absorber is plotted. Notice the radical difference between the graph of I_{12} for the water waves and the graph of P_{12} for the bullets. The difference is due to interference. Likewise, when we observe light waves, we also observe the intensity distribution, not the wave amplitude.

For quantum waves, we already know that the property that is proportional to the square of the wave amplitude is probability. We now need to find out what interference implies for the measurement of probabilities.

Let y_1 and y_2 be the amplitudes, or heights, of two probability waves representing indistinguishable particles measured at the same point in space. (In quantum theory, these amplitudes are generally complex quantities. For simplicity, here we assume they are real.) The sum of these two heights is simply $\psi = \psi_1 + \psi_2$, so the probability is $\psi^2 = (\psi_1 + \psi_2)^2 = \psi_1{}^2 + 2\psi_1\psi_2 + \psi_2{}^2$

This equation has a simple interpretation. The first term on the right is simply the probability that the first particle would appear if there were no interference from the second particle, and vice versa for the last term. Thus these two terms by themselves could represent the probabilities for classical particles like bullets, even though we do not ordinarily represent them by waves. If the middle term did not exist, this expression would then just represent the sum of two such classical probabilities. In Fig., it would represent the probability that a bullet came through either the first hole or the second hole and appeared at a particular point on the screen.

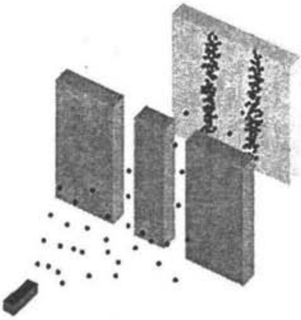


Fig. 17 Probability of Bullet from Hole.

The middle term on the right is called the interference term. This term appears only for wave phenomena and is responsible for destructive or constructive interference since it can be either negative or positive. If destructive interference is complete, the middle term completely cancels the other two terms (this will happen if $y_1 = -y_2$). Probability distributions for waves are completely different from those for bullets because of interference. The probability distribution for electrons has the same shape as the intensity distribution of the water waves shown in the middle figure because both distributions are derived from the square of algebraically summed wave amplitudes.

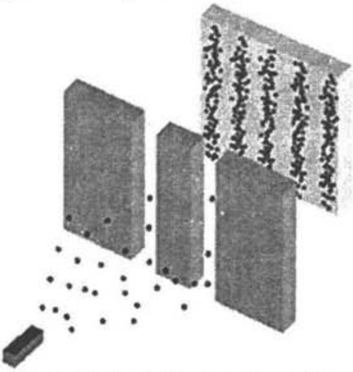


Fig. 18 Actual Electron Impacts.

We can now state an important conclusion from this discussion. Whenever we observe interference, it suggests the existence of real, external, objective waves rather than merely fictitious waves that are only tools for calculating probabilities of outcomes.

Remember that when we detect quantum waves, we detect particles. Since we are detecting particles, it may seem that the particle must come from one hole or the other, but that is incorrect. The particles that we detect do not come from the holes, they appear at the time of detection. Prior to detection, we have only probability waves.

What happens if we try to see whether we actually have electrons to the left of the detection screen, perhaps by shining a bright light on them between the holes and the detection screen, and looking for reflected light from these electrons? If the light is intense enough to see every electron this way before it is detected at the screen, the interference pattern is obliterated, and we see only the classical particle distribution shown in the top figure. Any measurement which actually manifests electrons to the left of the screen, such as viewing them under bright light, eliminates the probability wave which originally produced the interference pattern. After that we see only particle probability wave distributions.

Schrodinger's Cat Paradox

This thought experiment was originally created by Schrodinger in an attempt to show the possible absurdities if quantum theory were not confined to microscopic objects alone. Schrodinger thought the wave properties of the microworld could be transmitted to the macroworld if the former is coupled to the latter.

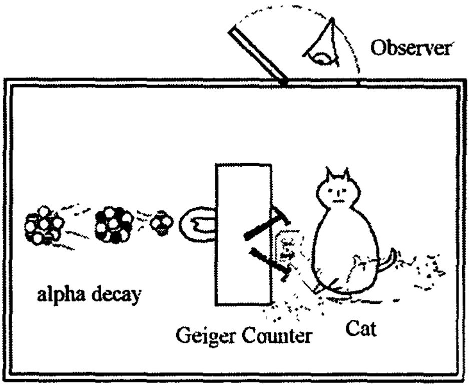


Fig. 19 Cat Paradox.

Imagine a closed box containing a single radioactive nucleus and a particle detector such as a Geiger counter. We assume this detector is designed to detect with certainty any particle that is emitted by the nucleus. The radioactive nucleus is microscopic and therefore can be described by quantum theory. Suppose the probability that the source will emit a particle in one minute is 1/2=50 per cent. The period of one minute is called the half-life of the source. Since the wavefunction of the nucleus is a solution to the Schrodinger equation and must describe all possibilities, after one minute it consists of a wave with two terms, one corresponding to a nucleus with one emitted particle, and one corresponding to a nucleus with no emitted particle, both measured at the same point in space:

$$\psi = \psi_1 \text{ (particle)} + \psi_2 \text{ (no particle)}$$

where, for simplicity, we again assume the wavefunctions are real rather than complex. Now, ψ_1^2 is the probability that a measurement would show that a particle was emitted, and ψ_2^2 is the probability that it would show that no particle was emitted.

The remaining items in the box are all macroscopic, but because they are nothing more than collections of microscopic particles (atoms and molecules) that obey quantum theory, we assume they also obey quantum theory.

If macroscopic objects do not obey quantum theory, we have no other theory to explain them. Classical physics is inadequate because it cannot explain the following experimental observations: Interference fringes have been directly produced with buckminsterfullerenes ('buckyballs') consisting of 60 carbon atoms and 48 fluorine atoms ($C_{60}F_{48,}$). Many much larger systems also show quantum effects. A superconducting quantum interference device (SQUID) containing millions of electrons was made to occupy Schrodinger's cat states. Ferromagnetism, superconductivity, and superfluidity all are quantum phenomena which occur in macroscopic systems.

Hence, we assume the Geiger counter can also be described by a wavefunction that is a solution to the Schrodinger equation. The combined system of nucleus and detector then must be described by a wavefunction that contains two terms, one describing a nucleus and a detector that has detected a particle, and one describing a nucleus and a detector that has not detected a particle:

$\psi = \psi_1$(detected particle) + ψ_2(no detected particle)

Both of these terms must necessarily be present, and the resulting state y is a superposition of these two states. Again, $\psi_1{}^2$ and $\psi_2{}^2$ are the probabilities that a measurement would show either of the two states.

Put into the box a vial of poison gas and connect it to the detector so that the gas is automatically released if the detector counts a particle. Now put into the box a live cat. We assume that the poison gas and cat can also be described by the Schrodinger equation. The final wavefunction contains two terms, one describing a detected particle, plus released gas and a dead cat; and one describing no detected particle, no released gas, and a live cat. Both terms must be present if quantum theory can be applied to the box's contents. The wavefunction must describe both a dead cat and a live cat:

$\psi = \psi_1$(detected particle, dead cat) + ψ_2(no detected particle, live cat)

After exactly one minute, you look into the box and see either a live cat or a dead one, but certainly not both! What is the explanation?

Until there is an observation, there is no cat, live or dead! There is only a wavefunction. The wavefunction merely tells us what possibilities will be presented to the observer when the box is opened. The observation itself manifests the reality of either a live cat or a dead cat (this is called reality).

Now we must ask why the observer him/her self is not included in the system described by the Schrodinger equation, so we put it in the following equation:

$\psi = \psi_1$(detected particle, observer sees dead cat) + ψ_2(no detected particle, observer sees live cat)

We know that the observer can observe only a live or a dead cat, not both. (The interference term $2\psi_1\psi_2$ does not contribute because it represents the observation of a live cat superimposed on a dead cat. Such an observation would be contrary to our experience and therefore cannot be allowed by the theory.) Hence, something about the observer cannot be described by the Schrodinger equation. What is this property? The one distinguishing property that is not described by quantum theory is consciousness. Hence, some physicists conclude that it must be consciousness which defines an observation.

Until now, this discussion has assumed that the observer but not the cat is conscious. But what if the cat is conscious? Then its own consciousness will define a continuous set of observations as long as it is alive. However, there is a 50 per cent probability that the poison gas will be released and will kill it within one minute. If that happens, its consciousness disappears. One could say that the cat's own consciousness killed it (but of course, without it, there would not have been a cat).

Live animals are not needed to show Schrodinger's paradox. The poison gas and cat can be omitted because the detector is assumed to be a macroscopic device that changes state upon detection of a microscopic particle. If the outside observer sees that the detector is in one state (by reading an indicator) prior to closing the box, and in either the same or a different state after reopening the box, the paradox is again demonstrated. Prior to reopening the box, the contents are in a superposition of two quantum waves. After opening the box, the observer sees that the detector is in one state or the other, not both.

Bell's Theorem, the Aspect-Gröblacher Experiments and the Nonlocality of Reality

One of the principles considered most sacred by Einstein and indeed by most physicists up until the 1980s is the principle of local causality, or locality for short. This principle (which comes from Einstein's theory of special relativity) states that no physical effect can be transmitted with a velocity faster than light. Also implied, but not always stated, is the principle that all physical effects must decrease as the distance between the source of the effect and the observer increases. In practice, this principle prohibits not only all instantaneous action-at-a-distance, but also any action-at-a-distance when the distances are so large that the longest-range known force that can transmit signals, the electromagnetic force, cannot feasibly produce the effect. If the particles of a system are assumed to be independent of each other except for physical effects that travel no faster than the velocity of light, the system is said to be local. This means, e.g., that if a measurement is made on one particle, the other particles cannot be affected before a local signal from the first particle can reach them.

In addition to locality, the other strongly held principle was the principle of objective reality. This principle states that there is a reality that exists whether or not it is observed. Prior to the discovery of quantum mechanics, this meant that this reality consisted of material particles or waves that always had definite physical properties, and which could become known either by making a measurement or by calculation using classical laws and a known initial state. For example, a particle always had a definite position and velocity prior to measurement, even though they may not have been known until a measure-ment or calculation was made. We call this strong objectivity. After the development of quantum mechanics, those who believe in an observer-created reality believe that only a wavefunction exists prior to an observation but this is still considered to be objectively real. However, its physical parameters, such as position and velocity, are indefinite until a measurement is made. This is called weak objectivity.

Weak objectivity was difficult enough to accept by some physicists, but quantum theory predicted something else that was even harder to accept that reality is nonlocal. This means that a measurement on one particle in a nonlocal system is correlated with a measurement on any of the other particles in the system even if no local signal passes from the first measurement to the second. For example, a measurement of the position of one particle in a nonlocal system is correlated with a position measurement of any of the other particles, indepe.ident of any local signals. A nonlocal system of particles is described by a wavefunction formed by a superposition of individual particle wavefunctions in such a way that all of the individual waves are locked together into a coherent whole. In such a coherent superposition, it is no longer possible to identify the individual particle components. The system behaves as a whole rather than as a collection of independent particles. We shall describe an example of a nonlocal system when we discuss Bell's theorem below.

Einstein could never accept a reality which was nonlocal or which was indefinite. His paper written with Podolsky and Rosen in 1935 was an attempt to use a thought experiment to show that, because quantum mechanics could not describe a reality which was both local and definite, the theory was incomplete.

Following the EPR paper, many physicists expended a great deal of effort in trying to devise theories that were complete, namely, theories that assumed that parameters like position and velocity are at all times definite even if they are unknown, and which at the same time gave results that agree with quantum theory. (These are called hidden variable theories, which by definition assume strong objectivity.) None of these theories found general acceptance because they were inelegant, complicated, and awkward to use, and the best-known version also turned out to be extremely nonlocal.

John Bell , brilliant, creative Northern Ireland physicist, devised a way to determine experimentally whether reality could be described by local hidden variable theories, and derived an inequality that was valid only if local hidden variable theories were valid. Furthermore, this inequality depended only on experimentally measured quantities, hence it was independent of any specific theory. Any violation of the inequality would prove that reality cannot be both strongly objective and local.

Many experiments were subsequently done to test his inequality, with the results that it was always violated, thus showing that if there is a strongly objective reality, it could not be local. In addition, the experiments always gave results that were consistent with the predictions of quantum theory. The best of these experiments were done by a group led by French physicist Alain Aspect in 1981-82. These results have far-reaching implications in the interpretation of quantum theory.

The Aspect experiments used pairs of photons, the two photons of each pair being emitted in opposite directions from a calcium source. These photon pairs had the property that the polarization directions (the vibration directions, which are always perpendicular to the propagation direction) of the two photons of a pair were always parallel to each other, but the polarization directions of different pairs were randomly distributed.

The two sides of the experiment were 12 metres apart. Each side had two detectors, to detect photons with two different polarization directions. Each detector separately recorded an equal number of photons for all polarization directions, showing that the photons were completely unpolarized. Now assume the detectors were wired to measure only coincidence counts, i.e., photons were recorded only if they were detected approximately simultaneously at A and B. Bell's inequality says that, if reality is local, a certain function S of these coincidence counts, measured for all four combinations of the two polarization angles A1, A2 and the two polarization angles B1, B2, must be between -2.0 and +2.0. The experiments yielded a value for S_{expt} of 2.70 ± 0.015. Thus Bell's inequality was violated.

Thus, the system in the Aspect experiments cannot be both strongly objective and local. This result is independent of whether or not quantum theory is valid.

These experiments could not distinguish between a reality that is not strongly objective but is local; one that is nonlocal but is strongly objective; and one that is neither strongly objective nor local. Furthermore, the measured value of the function S was always in agreement with the predictions of quantum theory (S_{QM} = 2.70 ± 0.05), which assumes that the photons are described by wavefunctions.

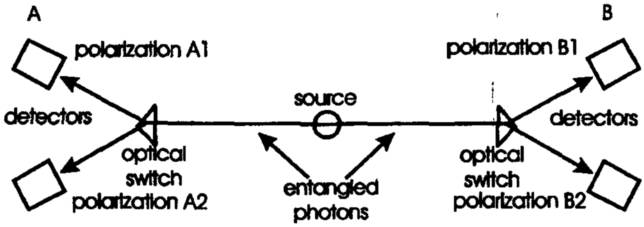


Fig. 20 Correlations between the Polarizations.

Bell's function F is a measure of the correlations between the polarizations (vibration directions) measured at the two sides A and B. The existence of correlations does not itself prove that reality is indefinite or nonlocal. In fact, correlations can exist between measurements at the two sides whether the photons are local and definite ('real' photons) or whether they are nonlocal and indefinite. If they are local and definite, correlations will exist if the two 'real' photons emitted by the source are individual particles that are polarized parallel (or perpendicular) to each other. If they are nonlocal and indefinite, correlations can exist if the system is described by a wavefunction that is a coherent superposition of the waves of the two photons (an 'entangled pair'). Because such a wavefunction represents a coherent whole rather than individual particles, it permits correlations that are greater than can exist with local, definite photons. That is why S is greater for entangled photons than for local, definite photons, and why the measured violation of Bell's inequality is consistent with photons described by quantum theory.

Now we must ask whether any class of hidden variable theories, which are all designed to be strongly objective, can be excluded by experiment. To help answer this question, an inequality similar to Bell's inequality was recently devised by Tony Leggett. The result disallowed the assumed class of hidden variable theories.

Groblacher *et al.* concluded that no hidden variable theory that is not counterintuitive (that is not bizarre) can describe reality. If so, then reality cannot be strongly objective, i.e., it can have no definite properties before measurement. The Aspect and Groblacher experiments taken together strongly imply that reality is both indefinite and nonlocal. This conclusion is independent of whether or not quantum theory is valid.

In a nonlocal system, a measurement made at one end of the system is correlated with a measurement made at the other end even if no local signal passes between the two. It might be thought that, because nonlocal correlations can exist between events occurring at two different points, observers at these two points could use these correlations to communicate instantaneously with each other in violation of Einstein's special theory of relativity. However, the nonlocality of quantum theory implies a correlation between data sets, not a transmission of information at greater than light velocities. Thus, the special theory is not violated. We can see this by realizing that the photons detected at either A or B alone occur completely randomly both in time and in polarization. Consequently, observer A sees no information in his data alone, and likewise with observer B. It is only by later comparing these two random sets of data that a correlation between the two sets can be discovered.

There can be strong correlations between two random sets that cannot be discovered by looking at one set alone. This is illustrated by the example of random stereograms which, when first viewed, look like near-random patterns of coloured dots. However, there are actually two separate near-random patterns present, and they are displaced from each other by a distance roughly equal to the spacing between a person's eyes. Thus, by looking at the pattern with the direction of the eyes nonconvergent as if looking some distance away, the two eyes see different patterns. The correlations between the patterns are discerned by the brain, and a three-dimensional image is seen.

3

Path Integrals in Quantum Mechanics

Huygen's Picture of Wave Propagation

If a point source of light is switched on, the wavefront is an expanding sphere centered at the source. Huygens suggested that this could be understood if at any instant in time each point on the wavefront was regarded as a source of secondary wavelets, and the new wavefront a moment later was to be regarded as built up from the sum of these wavelets. For a light shining continuously, this process just keeps repeating.

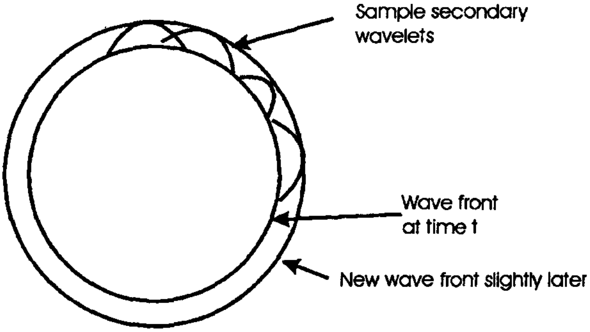


Fig. 1 Huygen's Picture of how a Spherical Wave Propagates: Each point on the Wave Front is a Source of Secondary Wavelets that Generate the new Wave Front.

What use is this idea? For one thing, it explains refraction the change in direction of a wavefront on entering a different medium, such as a ray of light going from air into glass.

If the light moves more slowly in the glass, velocity *v* instead of *c*, with $v < c$, then Huygen's picture explains Snell's Law, that the ratio of the sines of the angles to the normal of incident and transmitted beams is constant, and in fact is the ratio *c/v*. This is evident from Fig. 2 below: in the time the wavelet centered at A has propagated to *C*, that from *B* has reached *D*, the ratio of lengths *AC/ BD* being *c/v*. But the angles in Snell's Law are in fact the angles *ABC, BCD*, and those right-angled triangles have a common hypotenuse *BC*, from which the Law follows.

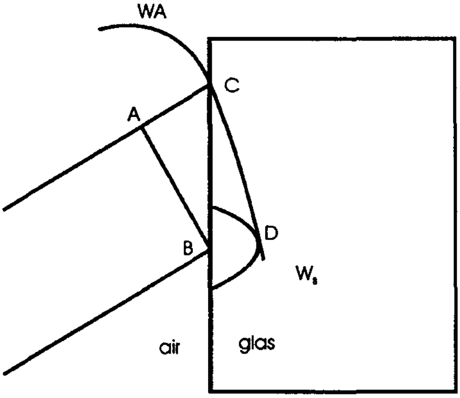


Fig. 2 Hugens' explanation of refraction: Showing two Wavelets from the Wavefront AB. W_B is slowed down compared with $W_{A/}$ Since it is Propagating in glass. This turns the Wave front through an Angle.

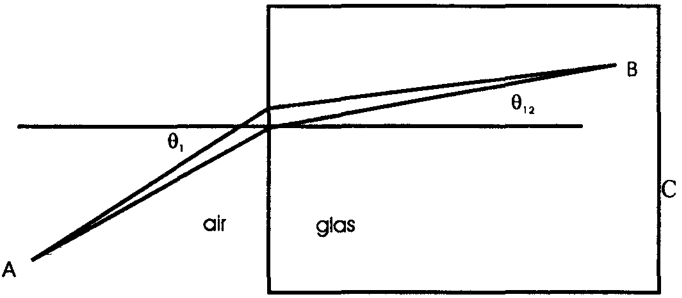


Fig. 3 Refraction of Wave.

Where the air meets the glass, the two rays, separated by a small distance CD = d along that interface, will look parallel:

Fermat's Principle of Least Time

We will now temporarily forget about the wave nature of light, and consider a narrow ray or beam of light shining from point *A* to point *B*, where we suppose *A* to be in air, *B* in glass. Fermat showed that the path of such a beam is given by the Principle of Least Time: a ray of light going from *A* to *B* by any other path would take longer. How can we see that? It's obvious that any deviation from a straight line path in air or in the glass is going to add to the time taken, but what about moving slightly the point at which the beam enters the glass?

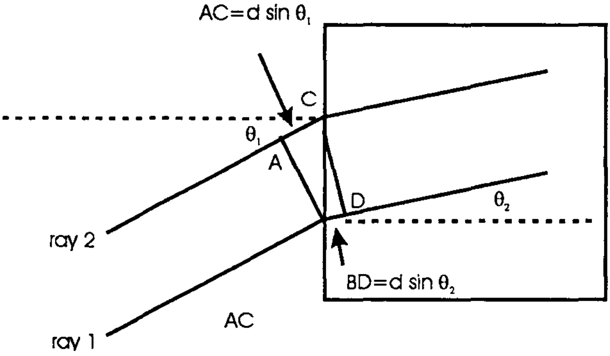


Fig. 4 : Magnified View of 2 rays Passing Through interface: ray 1 is the Minumum Time Path. Rays Encounter Interface Distance CB=d Aparts.

(Feynman gives a nice illustration: a lifeguard on a beach spots a swimmer in trouble some distance away, in a diagonal direction. He can run three times faster than he can swim. What is the quickest path to the swimmer?)

Moving the point of entry up a small distance *d*, the light has to travel an extra $d \sin\theta_1$ in air, but a distance less by $d \sin\theta_2$ in the glass, giving an extra travel time $\Delta t = d\sin\theta_1 / c - d\sin\theta_2 / v$. For the classical path, Snell's Law gives $\sin\theta_1 / \sin\theta_2 = n = c / v$, so $\Delta t = 0$ to first order. But if we look at a series of possible paths, each a small distance d away from the next at the point of crossing from air into glass, Δt becomes of order *d/c* away from the classical path.

Suppose now we imagine that the light actually travels along all these paths with about equal amplitude. (This actually is what Huygen's picture suggests: if we imagine the wavefront to generate secondary wavelets every picosecond, say, we can visualize the paths as zigzags with steps of length 3 mm.) What will be the total contribution of all these paths at *B?* Since the times along the paths are different, the signals along the different paths will arrive at *B* with different phases, and to get the total wave amplitude we must add a series of unit *2D* vectors, one from each path. (Representing the amplitude and phase of the wave by a complex number for convenience for a real wave, we can take the real part at the end.)

When we map out these unit *2D* vectors, we find that in the neighbourhood of the classical path, the phase varies little, but as we go away from it the phase spirals more and more rapidly, so those paths interfere amongst themselves destructively.

This is the explanation of Fermat's Principle only near the path of least time do paths stay approximately in phase with each other and add constructively. So this classical path rule has an underlying wave-phase explanation. In fact, the central role of phase in this analysis is sometimes emphasized by saying the light beam follows the path of stationary phase.

The Principle of Least Action

Confining our attention for the moment to the mechanics of a single nonrelativistic particle in a potential, with Lagrangian $L = T$ - V, the action S is defined by

$$s = \int_{t_1}^{t_2} L(x, x)\, dt,$$

Newton's Laws of Motion can be shown to be equivalent to the statement that a particle moving in the potential from A at t_1 to B at t_2 travels along the path that minimizes the action. This is called the *Principle of Least Action*: for example, the parabolic path followed by a ball thrown through the air minimizes the integral along the path of the action T–V where T is the ball's kinetic energy, V its gravitational potential energy (neglecting air resistance, of course). Note here that the initial and final times are fixed, so since we'll be summing over paths with different lengths, necessarily the particles speed will be different along the different paths. In other words, it will have different energies along the different paths.

With the advent of quantum mechanics, and the realization that any particle, including a thrown ball, has wave like properties, the rather mysterious Principle of Least Action looks a lot like Fermat's Principle of Least Time. Recall that Fermat's Principle works because the total phase along a path is the integrated time elapsed along the path, and for a path where that integral is stationary for small path variations, neighbouring paths add constructively, and no other sets of paths do. If the Principle of Least Action has a similar explanation, then the wave amplitude for a particle going along a path from *A* to *B* must have a phase equal to some constant times the action along that path. If this is the case, then the observed path followed will be just that of least action, for only near that path will the amplitudes add constructively, just as in Fermat's analysis of light rays.

Going from Classical Mechanics to Quantum Mechanics

Of course, if we write a phase factor for a path e^{icS} where *S* is the action for the path and *c* is some constant, *c* must necessarily have the dimensions of inverse action. Fortunately, there is a natural candidate for the constant *c*. The wave nature of matter arises from quantum mechanics, and the fundamental constant of quantum mechanics, Planck's constant, is in fact a unit of action. It turns out that the appropriate path phase factor is.

That the phase factor is $e^{is/k}$, rather than $e^{is/k}$, say, can be established by considering the double slit experiment for electrons.

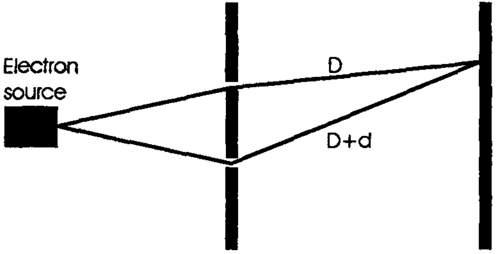


Fig. 5 Ddouble Slit Experiment for Electronsouble Slit Experiment for Electrons.

Suppose electrons from the top slit, Path I, go a distance *D* to the detector, those from the bottom slit, Path II, go *D* + *d*, with *d* << *D*. Then if the electrons have wavelength λ we know the phase difference at the detector is $2\pi d / \lambda$. To see this from our formula for summing over paths, on Path I the action $S = Et = \frac{1}{2}mv_1^2t$, and $v_1 = D/t$, so

$$S_1 = \frac{1}{2}mD^2/t.$$

For Path II, we must take $v_2 = (D + d)/t$. Keeping only terms of leading order in d/D, the action difference between the two paths

$$S_2 - S_1 = mDd/t$$

So the phase difference

$$\frac{S_2 - S_1}{\hbar} = \frac{mvd}{\hbar} = \frac{2\pi pd}{h} = \frac{2\pi d}{\lambda}.$$

This is the known correct result, and this fixes the constant multiplying the action/h in the expression for the path phase.

In quantum mechanics, such as the motion of an electron in an atom, we know that the particle does not follow a well-defined path, in contrast to classical mechanics. Where does the crossover to a well-defined path take place? Taking the simplest possible case of a free particle (no potential) of mass m moving at speed v, the action along a straight line path taking time t from A to B is $\frac{1}{2}mv^2t$. If this action is of order Planck's constant h, then the phase factor will not oscillate violently on moving to different paths, and a range of paths will contribute. In other words, quantum rather than classical behaviour dominates when $\frac{1}{2}mv^2t$ is of order h. But vt is the path length L, and mv/h is the wavelength λ, so we conclude that we must use quantum mechanics when the wavelength h/p is significant compared with the path length. Interference sets in when the difference in path actions is of order h, so in the atomic regime many paths must be included.

Feynman (in Feynman and Hibbs) gives a nice picture to help think about summing over paths. He begins with the double slit experiment for an electron. We suppose the electron is emitted from some source A on the left, and we look for it at a point B on a screen to the right. In the middle is a thin opaque barrier with the familiar two slits. Evidently, to find the amplitude for the electron to reach B we sum over two paths. Now suppose we add *another* two-slit barrier. We have to sum over four paths. Now add another. Next, replace the two slits in each barrier by several slits. We must sum over a multitude of paths! Finally, increase the number of barriers to some large number N, and at the same time increase the number of slits to the point that there are no barriers left. We are left with a sum over all possible paths through space from A to B, multiplying each path by the appropriate action phase factor.

In fact, the sum over paths is even more daunting than this picture suggests. All the paths going through these many slitted barriers are progressing in a forward direction, from A towards B. Actually, if we're summing over all paths, we should be including the possibility of paths zigzagging backwards and forwards as well, eventually arriving at B. We shall soon see how to deal systematically with all possible paths.

The Free Electron Propagator

As a warm up exercise, we consider an electron confined to one dimension, with no potential present, moving from at time 0 to x at time t. (This is speaking loosely we mean, as explained previously, that the initial state of the electron is a normalizable state, such as a Gaussian, concentrated closely at. The propagator then represents the probability amplitude, that is, the wave function, at point x after the given time t.) The propagator is given by

$$|\psi(x,t)\rangle = U(t)|\psi(x,t=0)\rangle,$$

or, in Schrodinger wave function notation,

$$\psi(x,t) = \int U\left(x,t;x',t=0\right)\psi\left(x',t=0\right)dx'.$$

It is clear that for this to make sense, as

$$t \to 0,\ \ U\left(x,t;x',0\right) \to \delta\left(x-x'\right)$$

$$\langle x|U(t,0)|x'\rangle = \int_{-\infty}^{\infty} e^{ikk^2 t/2m}\frac{dk}{2\pi}\langle x|k\rangle\langle k|x'\rangle$$

$$= \int_{-\infty}^{\infty} e^{-ikk^2 t/2m}\frac{dk}{2\pi}e^{-ik(x-x')}$$

$$= \sqrt{\frac{m}{2\pi\hbar it}}e^{im(x-x')^2/2kt}$$

Now let us think about the sum over paths. Let us *assume* that the classical path dominates, and that only paths in its neighbourhood contribute, and all the other paths do is multiply the effect of the single classical path by some function of time. (This arbitrary seeming assumption depends heavily on knowing the answer in advance. The classical path, of course, corresponds to motion from *x*' to *x* at a constant speed $v = (x - x')/t$. The action along this path is therefore *Et*, where *E* is the classical energy ½ mv^2, giving

$$U(x,t;x',0) = A'e^{im(x-x')^2/2kt}$$

This gives the correct exponential term! The prefactor *AÁ* can be determined from the requirement that as *t* goes to zero, *U* must approach a δ-function. This gives the correct prefactor, identical to the one found previously.

However, we have been lucky in more interesting situations, the classical path doesn't give all the information, and we really must address the issue of integrating over all paths.

Proving that the Sum-Over-Paths Definition of the Propagàtor is Equivalent to the Sum-Over-Eigenfunctions Definition

The first step is to construct a practical method of summing over paths. Let us begin with a particle in one dimension going from *x'* at time *t'* to *x* at time *t*. The paths can be enumerated in a crude way, reminiscent of Riemann integration: divide the time interval *t'* to *t* into *N* equal intervals each of duration

$$\varepsilon, t_1 = t_0 + \varepsilon,\ t_2 = t_0 + 2\varepsilon,\ \ldots.,\ t_N = t \text{ so on.}$$

Next, define a particular path from *x* to *x*' by specifying the position of the particle at each of the intermediate times, that is to say, it is at x_1 at time t_1, x_2 at time t_2 and so on. Then, simplify the path by putting in straight line bits connecting x_0 to x_1, x_1 to x_2, etc. The justification is that in the limit of ε going to zero, taken at the end, this becomes a true representation of the path.

The next step is to sum over all possible paths with a factor $e^{iS/k}$ for each one. The sum is accomplished by integrating over all possible values of the intermediate positions $x_1, x_2, \ldots x_{N-1}$ and then taking *N* to infinity.

The action on the zigzag path is

$$S = \int_{t'}^{t} dt(\frac{1}{2}mx^{-2} - V(x)) \rightarrow \sum_i \left[\frac{m(x_{i+1} - x_i)^2}{2\varepsilon} - \varepsilon V\left(\frac{x_{i+1} + x}{2} \right) \right]$$

We define the ‘integral over paths’ written by $\int D\left[x(t)\right]$

$$\lim_{\substack{r\to 0\\ N\to\infty}} \frac{1}{B(\varepsilon)}\int_{-\infty}^{\infty}\int\ldots\int\frac{dx_1}{B(\varepsilon)}\cdots\frac{dx_{N-1}}{B(\varepsilon)}$$

where we haven’t yet figured out what the overall weighting factor $B(\varepsilon)$ is going to be. (It is standard convention to have that extra outside.)

To summarize: the propagator $U(x,t;x', t')$ is the contribution to the wave function at x at time t from that at x' at the earlier time t'.

Consequently, $U(x,t;x',t')$ regarded as a function of x, t is, in fact, nothing but the Schrodinger wave function $\psi(x,t)$, and therefore must *satisfy Schrodinger's equation*

$$i\hbar\frac{\partial}{\partial t}U(x,t;x',t')\left(-\frac{\hbar}{2m}\frac{\partial^2}{\partial x^2}+V(x)\right)U(x,t;x',t').$$

We shall now show that defining $U\left(x,t;x',t'\right)$ as a sum over paths, it does in fact satisfy Schrodinger’s equation, and furthermore goes to a-function as time goes to zero.

$$U(x,t;x',t')=\int D\left[x(t)\right]e^{iS\left[x(t)\right]/k}$$

$$=\lim_{\substack{z\to 0\\ N\to\infty}} \frac{1}{B(\varepsilon)}\int_{-\infty}^{\infty}\int\ldots\int\frac{dx_1}{B(\varepsilon)}\cdots\frac{dx_{N-1}}{B(\varepsilon)}e^{is(x_1,\ldots,x_{N-1})/k}$$

We shall establish this equivalence by proving that it satisfies the same differential equation. It clearly has the same initial value t' as and t coincide, it goes to $\delta\ (x\text{-}x')$ in both representations.

To differentiate $U\,(x,t;x',t')$ with respect to t, we isolate the integral over the last path variable, $x_{N\text{-}1}$:

$$U(x,t;x',t')=\int\frac{dx_{N-1}}{B(\varepsilon)}e^{\left[\frac{im(x-x_{N-1})^2}{2k\varepsilon}-\frac{\imath}{k}\tau V\left(\frac{x+x_{N-1}}{2}\right)\right]}U\left(x_{N-1},t-\varepsilon,x',t'\right)$$

Now in the limit ε going to zero, almost all the contribution to this integral must come from close to the point of stationary phase, that is, $x_{N\text{-}1}$ = x. In that limit, we can take $U\left(x_{N\text{-}1},t\text{-}\varepsilon,x',t'\right)$ to be a slowly varying function of $x_{N\text{-}1}$, and replace it by the leading terms in a Taylor expansion about x, so

$$U(x,t;x',t')=$$

$$\int\frac{dx_{N-1}}{B(\varepsilon)}e^{\frac{im(x-x_{N-i})^2}{2kt}}\left(1-\frac{i}{\hbar}-\varepsilon V\left(\frac{x+x_{N-1}}{2}\right)\right)$$
$$\left(U(x,t-\varepsilon)+(x_{N-1}-x)\frac{\partial U}{\partial x}+\frac{(x_{N-1}-x)^2}{2}\frac{\partial^2}{\partial x^2}\right)$$

The x_{N-1} dependence in the potential V can be neglected in leading order that leaves standard Gaussian integrals, and

$$U(x,t;x',t')=\frac{1}{B(\varepsilon)}\sqrt{\frac{2\pi\hbar\varepsilon}{-im}}\left(1-\frac{i\varepsilon}{\hbar}V(x)+\frac{i\varepsilon\hbar}{2m}\frac{\partial^2}{\partial x^2}\right)U(x,t-\varepsilon;x't').$$

Taking the limit of ε going to zero fixes our unknown normalizing factor,

$$B(\varepsilon)=\sqrt{\frac{2\pi\hbar\varepsilon}{-im}}$$

giving

$$i\hbar\frac{\partial}{\partial t}U(x,t;x',t')=\left(-\frac{\hbar^2}{2m}\frac{\partial^2}{\partial x^2}+V(x)\right)U(x,t'x',t'),$$

thus establishing that the propagator derived from the sum over paths obeys Schrodinger's equation, and consequently gives the same physics as the conventional approach.

The Path Integral for the Free Particle

The required correspondence to the Schrodinger equation result fixes the unknown normalizing factor, as we've just established. This means we are now in a position to evaluate the sum over paths explicitly, at least in the free particle case, and confirm the somewhat handwaving result given above.

The sum over paths is

$$U(x,t;x',t')=\int D[x(t)]e^{iS[x(t)]/k}=\lim_{\substack{\tau\to 0\\ N\to\infty}}\frac{1}{B(\varepsilon)}\int_{-\infty}^{\infty}\int\ldots\int\frac{dx_1}{(\varepsilon)}$$

$$\frac{dx_{N-1}}{B(\varepsilon)} e^{i\sum_i \frac{m(x_{i+i}-x_i)^2}{2kt}}$$

Let us consider the sum for small but finite ε. In particular, we'll divide up the interval first into halves, then quarters, and so on, into 2^n small intervals. The reason for this choice will become clear.

Now, we'll integrate over half the paths: those for *i* odd, leaving the even x_i values fixed for the moment. The integrals are of the form

$$\int_{-\infty}^{\infty} dy e^{(ia/2)\left[(x-y)^2+(y-z)^2\right]} = e^{(ia/2)\left(x^2+z^2\right)} \int_{-\infty}^{\infty} dy e^{iay^2 - iay(x+z)}$$

$$= e^{(ia/2)\left(x^2+z^2\right)} \sqrt{\frac{\pi}{-ia}} e^{(-ia/4)(x+z)^2} = \sqrt{\frac{\pi}{-ia}} e^{(ia/4)(x-z)^2}$$

using the standard result $\int_{-\infty}^{\infty} dx e^{-ax^2+bx} = \sqrt{\frac{\pi}{a}} e^{b^2/4a}$.

Now put in the value $a = m/\hbar\varepsilon$: the factor $\sqrt{\frac{\pi}{-ia}} = \sqrt{\frac{\pi\hbar\varepsilon}{-im}}$ cancels the normalization factor $B(\varepsilon)\sqrt{\frac{2\pi\hbar\varepsilon}{-im}}$ except for the factor of 2 inside the square root. But we need that factor of 2, because we're left with an integral over the remaining even numbered paths exactly like the one before except that the time interval has doubled, both in the normalization factor and in the exponent, $\varepsilon \to 2\varepsilon$.

So we're back where we started. We can now repeat the process, halving the number of paths again, then again, until finally we have the same expression but with only the fixed endpoints appearing.

4

Angular Momentum

Translation and Rotation Operators

As a warm up to analyzing how a wave function transforms under rotation, we review the effect of *linear translation* on a single particle wave function $\psi(x)$. We have already seen an example of this: the coherent states of a simple harmonic oscillator discussed earlier were (at $t = 0$) identical to the ground state except that they were centered at some point displaced from the origin. In fact, the operator creating such a state from the ground state is a translation operator.

The *translation operator* $T(a)$ is defined at that operator which when it acts on a wave function ket $|\psi(x)\rangle$ gives the ket corresponding to that wave function moved over by a, that is,

$$T(a)|\psi(x)\rangle = |\psi(x-a)\rangle,$$

so, for example, if $\psi(x)$ is a wave function centered at the origin, $T(a)$ moves it to be centered at the point a.

We have written the wave function as a ket here to emphasize the parallels between this operation and some later ones, but it is simpler at this point to just work with the wave function as a function, so we will drop the ket bracket for now. The form of $T(a)$ as an operator on a function is made evident by rewriting the Taylor series in operator form:

$$\psi(x-a) = \psi(x) - a\frac{d}{dx}\psi(x) + \frac{a^2}{2!}\frac{d^2}{dx^2}\psi(x) - \ldots$$

$$= e^{-a\frac{d}{dx}}\psi(x)$$

$$= T(a)\psi(x)$$

Now for the quantum connection: the differential operator appearing in the exponential is in quantum mechanics proportional to the momentum operator ($p = ihd / dx$) so the translation operator

$$T(a) = e^{-\iota\tau p/\hbar}$$

An important special case is that of an infinitesimal translation,

$$T(\varepsilon) = e^{-\iota\tau p/\hbar} = 1 - i\varepsilon p/\hbar$$

The momentum p is said to be the *generator* of the translation.

$$\mathrm{T}(\varepsilon)|\mathrm{x}_0\rangle = |\mathrm{x}_0 \mp \varepsilon\rangle$$

Here $|x\rangle$ denotes a delta-function type wave function centered at x. It might be better if he had written $T(\varepsilon)\partial(x–x_0) = \partial(x - x_0 - \varepsilon)$, then we would see right away that this translates into the wave function transformation $\mathrm{T}(\varepsilon)\partial(\mathrm{x} - \mathrm{x}_0) = \partial(\mathrm{x} - \mathrm{x}_0 - \varepsilon)$, the sign of now obviously consistent with our usage above.

It is important to be clear about whether the *system* is being translated by a, as we have done above or whether, alternately, the *coordinate axes* are being translated by a, that latter would result in the *opposite* change in the wave function. Translating the coordinate axes, along with the apparatus and any external fields by -a relative to the wave function would of course give the same physics as translating the wave function by $+a$. In fact, these two equivalent operations are analogous to the time development of a wave function being described either by a Schrodinger picture, in which the bras and kets change in time, but not the operators, and the Heisenberg picture in which the operators develop but the bras and kets do not change. To pursue this analogy a little further, in the 'Heisenberg' case

$$x \to T^{-1}(\varepsilon)xT(\varepsilon) = e^{\iota\tau p/k}x^{-e\iota\tau p/k} = x + i\varepsilon[p,x]/\hbar = x + \varepsilon$$

and π is unchanged since it commutes with the operator. So there are two possible ways to deal with translations: transform the bras and kets, *or* transform the operators. We shall almost always leave the operators alone, and transform the bras and kets.

We have established that *the momentum operator is the generator of spatial translations* (the generalization to three dimensions is trivial). We know from earlier work that the Hamiltonian is the generator of *time* translations, by which we mean

$$\psi(t+a)=e^{-\imath H_2/k}\psi(t).$$

It is tempting to conclude that the *angular momentum* must be the operator generating *rotations* of the system, and, in fact, it is easy to check that this is correct. Let us consider an infinitesimal rotation $\delta\vec{\theta}$ about some axis through the origin (the infinitesimal vector being in the direction of the axis). A wavefunction $\psi(r)$ initially localized at $\vec{r}_0$ will shift to be localized at $\vec{r}_0+\delta\vec{r}_0$, where $\delta\vec{r}_0=\delta\vec{\theta}\times\vec{r}_0$ So, how does a wave function transform under this small rotation? Just as for the translation case, $\psi(\vec{r})\to\psi(\vec{r}-\delta\vec{r})$. If you don't understand the minus sign, reread the discussion on translations and the sign of ε.

Thus

$$\psi(\vec{r})\to\psi(\vec{r})-\frac{i}{\hbar}\delta\,\vec{r}\,\vec{\hat{p}}\,\psi(\vec{r})$$

to first order in the infinitesimal quantity, so the rotation operator

$$R\left(\delta\vec{\theta}\right)\psi(\vec{r})=\left(1-\frac{i}{\hbar}\delta\vec{\theta}\times\vec{r}.\hat{\vec{p}}\right)\psi(\vec{r})$$

$$\left(1-\frac{i}{\hbar}\delta\vec{\theta}.\vec{r}\times\hat{\vec{p}}\right)\psi(\vec{r})$$

$$=\left(1-\frac{i}{\hbar}\delta\vec{\theta}.\hat{\vec{L}}\right)\psi(\vec{r}).$$

If we write this as

$$R\left(\delta\vec{\theta}\right)\psi(\vec{r})=e^{-\frac{i}{k}\delta\vec{\theta}\,\hat{\vec{L}}}\psi(\vec{r})$$

it is clear that a finite rotation is given by multiplying together a large number of these operators, which just amounts to replacing $\delta\vec{\theta}$ by $\vec{\theta}$ in the exponential. Another way of going from the infinitesimal rotation to a full rotation is to use the identity

$$\lim_{N\to\infty}\left(1+\frac{A\theta}{N}\right)^{N}=eA\theta$$

which is clearly valid even if A is an operator.

We have therefore established that the orbital angular momentum operator $\hat{\vec{L}}$ is the generator of spatial rotations, by which we mean that if we rotate our apparatus, and the wave function with it, the appropriately transformed wave function is generated by the action of $R(\vec{\theta})$ on the original wave function. It is perhaps worth giving an explicit example: suppose we rotate the system, and therefore the wave function, through an infinitesimal angle $\delta\theta_z$ about the z-axis. Denote the rotated wave function by $\psi_{\text{rot}}(x, y)$. Then

$$\psi_{rot}(x,y)=\left(1-\frac{i}{\hbar}(\delta\theta_z)L_z\right)\psi(x,y)$$

$$\left(1-\frac{i}{\hbar}(\delta\theta_z)\left(-i\hbar\left(x\frac{d}{dy}-y\frac{d}{dx}\right)\right)\right)\psi(x,y)$$

$$\left(1-(\delta\theta_z)\left(\left(x\frac{d}{dy}-y\frac{d}{dx}\right)\right)\right)\psi(x,y)$$

$$=\psi\left(x+(\delta\theta_z)y,y-(\delta\theta_z)x\right)$$

That is to say, the value of the new wave function at (x, y) is the value of the old wave function at the point which was rotated into (x, y).

Quantum Generalization of the Rotation Operator

However, it has long been known that in quantum mechanics, orbital angular momentum is *not* the whole story. Particles like the electron are found experimentally to have an internal angular momentum, called spin. In contrast to the spin of an ordinary macroscopic object like a spinning top, the electron's spin is *not* just the sum of orbital angular momenta of internal parts, and any attempt to understand it in that way leads to contradictions.

To take account of this new kind of angular momentum, we generalize the orbital angular momentum $\hat{\vec{L}}$ to an operator $\hat{\vec{J}}$ which is *defined* as the generator of rotations on *any* wave function, including possible spin components, so

$$R\left(\vec{\theta}\right)\psi\left(\vec{r}\right)=e^{-\frac{i}{\hbar}\delta\vec{\theta}\cdot\vec{J}}\psi\left(\vec{r}\right).$$

This is of course identical to the equation we found for *L*, but there we derived if from the quantum angular momentum operator including the momentum components written as differentials. But up to this point $\psi(\vec{r})$ has just been a complex valued function of position. From now on, the wave function at a point can have several components, so it is in some vector space, and the rotation operator will operate in this space as well as being a differential operator with respect to position. For example, the wave function could be a vector at each point, so rotation of the system could rotate this vector as well as moving it to a different $\vec{r}$.

To summarize: $\psi\left(\vec{r}\right)$ is in general an n-component function at each point in space, $R(\delta\vec{\theta})$ is an n × n matrix in the component space, and the above equation is the *definition* of *J*. Starting from this definition, we will find *J*'s properties.

The first point to make is that in contrast to translations, rotations do not commute even for a classical system. Rotating a book through $\pi/2$ first about the *z*-axis then about the *x*-axis leaves it in a different orientation from that obtained by rotating from the same starting position first $\pi/2$ about the *x*-axis then $\pi/2$ about the *z*-axis. Even small rotations do not commute, although the commutator is second order. Since the *R*-operators are representations of rotations, they will reflect this commutativity structure, and we can see just how they do that by considering ordinary classical rotations of a real vector in three-dimensional space.

$$R_x(\theta)\begin{pmatrix}1 & 0 & 0\\ 0 & \cos\theta & -\sin\theta\\ 0 & \sin\theta & \cos\theta\end{pmatrix}, R_y(\theta)=\begin{pmatrix}\cos\theta & 0 & \sin\theta\\ 0 & 1 & 0\\ -\sin\theta & 0 & \cos\theta\end{pmatrix},$$

$$R_z(\theta)\begin{pmatrix}\cos\theta & -\sin\theta & 0\\ \sin\theta & \cos\theta & 0\\ 0 & 0 & 1\end{pmatrix}$$

The matrices rotating a vector θ by about the x, y and z axes are in the limit of rotations about infinitesimal angles (ignoring higher order terms),

$$R_x(\varepsilon)=1+\varepsilon\begin{pmatrix}0&0&0\\0&0&-1\\0&1&0\end{pmatrix},R_y(\varepsilon)=1+\varepsilon\begin{pmatrix}0&0&0\\1&0&0\\-1&0&0\end{pmatrix},$$

$$R_z(\varepsilon)=1+\varepsilon\begin{pmatrix}0&-1&0\\1&0&0\\0&0&0\end{pmatrix}$$

It is easy to check that

$$[R_x(\varepsilon),R_y(\varepsilon)]=\varepsilon^2\begin{pmatrix}0&-1&0\\1&0&0\\0&0&0\end{pmatrix},\ .$$

which can also be written

$$R_x(\varepsilon),R_y(\varepsilon)=R_z(\varepsilon^2)R_y(\varepsilon)R_x(\varepsilon).$$

The rotation operators on quantum mechanical kets must, like all rotations, follow this same pattern, that is, we must have

$$\left(\left(1-\frac{i}{\hbar}-\varepsilon J_x\right)\left(1-\frac{i}{\hbar}\varepsilon j_y\right)\left(1-\frac{i}{\hbar}\varepsilon^2 J_z\right)\left(1-\frac{i}{\hbar}\varepsilon J_y\right)\left(1-\frac{i}{\hbar}\varepsilon J_x\right)\right)|\psi\rangle=0$$

where we have used the definition of the infinitesimal rotation operator on kets, $R\left(\delta\vec{\theta}\right)\psi(\vec{r})=e^{-\frac{i}{k}\delta\vec{\theta}\,\hat{\vec{J}}}\psi(\vec{r})$. The zeroth and first-order terms in e all cancel, the second-order term gives $\left[J_x,J_y\right]=ihJ_z$. The general statement is:

$$\left[J_i,J_j\right]=ih\varepsilon_{ijk}J_k$$

This is one of the most important formulas in quantum mechanics.

Consequences of the Commutation Relations

The commutation formula $\left[J_i,J_j\right]=ih\varepsilon_{ijk}J_k$ which is, after all, a straightforward extension of the result for ordinary classical rotations, has surprisingly far-reaching consequences: it leads directly to the directional quantization of spin and angular momentum observed in atoms subject to a magnetic field.

It is by now very clear that in quantum mechanical systems such as atoms the total angular momentum, and also the component of angular momentum in a given direction, can only take certain values. Let us try to construct a basis set of angular momentum states for a given system: a complete set of kets corresponding to all allowed values of the angular momentum. Now, angular momentum is a *vector* quantity: it has magnitude and direction. Let's begin with the magnitude, the natural parameter is the length squared:

$$J^2 = J_x^2 + J_y^2 + J_z^2 .$$

Now we must specify direction but here we run into a problem. J_x, J_y and J_z are all mutually non-commuting, so we cannot construct a set of common eigenkets of any two of them, which we would need for a precise specification of direction. They *do* all commute with J^2, since it is spherically symmetric and therefore cannot be affected by any rotation (and, it's easy to check this commutation explicitly).

The bottom line, then, is that in attempting to construct eigenkets describing the different possible angular momentum states of a quantum system, the best we can do is to find the common eigenkets of J^2 and *one* direction, say J_z. The commutation relations do not allow us to be more precise about direction, analogous to the Uncertainty Principle for position and momentum, which also comes from noncommutativity of the relevant operators.

We conclude that the appropriate angular momentum basis is the set of common eigenkets of the commuting Hermitian matrices J^2, J_z:

$$J^2 \left|a,b\right\rangle = a\left|a,b\right\rangle$$

$$J_z \left|a,b\right\rangle = b\left|a,b\right\rangle$$

Ladder Operators

The sets of allowed eigenvalues a,b can be found using the 'ladder operator' trick previously discussed for the simple harmonic oscillator. It turns out

$$J_\pm = J_x \pm ij_y$$

are closely analogous to the simple harmonic oscillator raising and lowering operators $a^\dagger$ and a.

J_+ and J_- have commutation relations with J_z:

$$[J_z, J_\pm] \pm \hbar J_\pm$$

and they of course *commute* with J^2, as do J_z, J_x and J_y.

Therefore, $J_\pm$ operating on $|a,b\rangle$ cannot affect the value of a. But they *do* change the value of b:

$$\begin{aligned}
&J_z, J_\pm |a,b\rangle = [J_z, K_\pm]|a,b\rangle + J_\pm J_z |a,b\rangle \\
&= \pm\hbar J_\pm |a,b\rangle + bJ_\pm |a,b\rangle \\
&= (b \pm \hbar) J_\pm |a,b\rangle
\end{aligned}$$

so if $|a, b>$ is an eigenket of J_z with eigenvalue b, $J_\pm |a, b>$ is either zero or an eigenket of J_z with eigenvalue $b \pm h$, that is, $J_\pm |a, b\rangle = C_\pm|a, b \pm h\rangle$ where $C_\pm(a,b)$ is a normalization constant, taking the initial $|a, b\rangle$ to be normalized. Just as with the simple harmonic oscillator, we have to find these normalization constants in order to compute matrix elements. All the physics is in the matrix elements.

The squared norm of $J_\pm |a, b\rangle$

$$\left\| J_\pm |a,h\rangle \right\|^2 = |a,b\rangle |J_\pm^\dagger J_\pm| |a,b\rangle = |a,b\rangle |J_\mp J_\pm| |a,b\rangle$$

and $J_\mp J_\pm = (J_x \mp ij_y)(J_x' \pm ij_y) = J_x^2 + J_y^2 \pm i\left[J_x, J_y\right]$

$$= J^2 - J_x^2 \mp \hbar J_z$$

from which

$$\left\| J_\pm |a,b\rangle \right\|^2 = \langle a,b| J^2 - J_x^2 \mp \hbar J |a,b\rangle = a - b^2 \mp \hbar b,$$

recalling that $\langle a,b|a,b\rangle = 1$.

Now a, being the eigenvalue of a sum of squares of Hermitian operators, is necessarily nonnegative, and b is real. Hence for a given a, b is *bounded*: there must be a $b_{\max}$ and a (negative or zero) $b_{\min}$. But this must mean that

$$\left\| J_+ |a, b_{\max}\rangle \right\|^2 = a - b_{\max}^2 - \hbar b_{\max} = 0\text{T}$$

Note that for a given a, b_{max} and b_{min} are determined uniquely there cannot be two kets with the same a but different b annihilated by J_+. It also follows immediately that $a = b_{max}(b_{max} + h)$ and $b_{min} = -b_{max}$ Furthermore, we know that if we keep operating on $|a, b_{min}>$ with J_+, we generate a sequence of kets with J_z eigenvalues $b_{min} + h$, $b_{min} + 3h$, $b_{min} + 3h$,.... This series must terminate, and the only possible way for that to happen is for b_{max} to be equal to $b_{min} + nh$ with n an integer, from which it follows that b_{max} is either an integer or half an odd integer times $\hbar$.

At this point, we switch to the standard notation. We have established that the eigenvalues of J_z form a finite ladder, spacing $\hbar$. We write them as, and j is used to denote the maximum value of m, so the eigenvalue of $J^2, a = j(j+1)\hbar^2$. Both j and m will be integers or half odd integers, but the *spacing* of the ladder of m values is always unity. Although we have been writing with we shall henceforth follow convention and write $|j,m\rangle$.

The operators $\vec{J}^2$, J_z have a common set of orthonormal eigenkets $|j,m\rangle$,

$$\vec{J}^2 |j,m\rangle = j(j+1)\hbar^2 |j,m\rangle$$

$$J_z |j,m\rangle = m\hbar |j,m\rangle$$

$$\langle j,m|j,m\rangle = 1$$

where j, m are integers or half integers. The allowed quantum numbers m form a ladder with step spacing unity, the maximum value of m is j, the minimum value is $-j$.

Normalizing J_+ and J_-

It is now straightforward to compute the normalization factors needed to find matrix elements:

$$\left\| J_\mp |j,m\rangle \right\|^2 = \langle j,m| J^2 - J_z^2 \mp \hbar J_z |j,m\rangle$$

$$= \left(j(j+1)\right)\hbar^2 - m(m \pm 1)\hbar^2)\langle j,m|j,m\rangle,$$

and, so

$$J_+ |j,m\rangle = \sqrt{j(j+1) - m(m+1)}\,\hbar |j,m+1\rangle$$

$$J_- |j,m\rangle = \sqrt{j(j+1) - m(m+1)}\,\hbar |j,m-1\rangle$$

With these formulas, and the base set of normalized eigenkets $|j, m\rangle$, we are in a position to construct explicit matrix representations of the angular momentum algebra for any integer or half integer value of angular momentum j.

Historical Note

The use of *m* to denote the component of angular momentum in one direction came about because a Bohr-type electron in orbit is a current loop, with a magnetic moment parallel to its angular momentum, so the *m* measured the component of magnetic moment in a chosen direction, usually along an external magnetic field, and *m* is often called the magnetic quantum number.

5

Orbital Eigenfunctions: 2-D and 3-D

Orbital Angular Momentum Eigenfunctions

We know that the operators $\vec{J}^2, J_z$ have a common set of eigenkets $|j,m\rangle$, $\vec{J}^2|j,m\rangle = j(j+1)\hbar^2|j,m\rangle$, $J_z|j,m\rangle = m\hbar|j,m\rangle$ where *j*, *m* are integers or half odd integers, and we found the matrix elements of (and hence those of J_x, J_y) between these eigenkets. This purely formal structure, therefore, nails down the allowed values of total angular momentum and of any measured component. But there are other things we need to know: for example, how is an electron in a particular angular momentum state in an atom affected by an external field? To compute that, we need to know the wave function.

If a system has spherical symmetry, such as an electron in the Coulomb field of a hydrogen nucleus, then the Hamiltonian *H* and the operators $\vec{J}^2, J_z$ have a common set of eigenkets $|E,j,m\rangle$. The spherically symmetric Hamiltonian is unchanged by rotation, so must commute with any rotation operator, $\left[H,\vec{J}^2\right]=0$ and $\left[H,j_z\right]=0$. Recall that commuting Hermitian operators can be diagonalized simultaneously and therefore have a common set of eigenkets.

Fortunately, many systems of interest do have spherical symmetry, at least to a good approximation, the basic example of course being the hydrogen atom, so the natural set of basis states is the common eigenkets of energy and angular momentum. It turns out that even when the spherical symmetry is broken, the angular momentum eigenkets may still be a useful starting point, with the symmetry breaking treated using perturbation theory.

Two-Dimensional Models

As a warm-up exercise for the complications of the three-dimensional spherically symmetric model, it is worth analyzing a two-dimensional *circularly* symmetric model, that is,

$$H\psi(x,y)=-\frac{\hbar^2}{2M}\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}\right)\psi(x,y)+V\left(\sqrt{x^2+y^2}\right)\psi(x,y)$$
$$=E\psi(x,y).$$

(In this section, we'll denote the particle mass by *M*, to avoid confusion with the angular momentum quantum number *m* + but be warned you are often going to find *m* used for both in the same discussion!)

The two-dimensional angular momentum operator is

$$L=\vec{r}\times\vec{p}=xp_y-yp_x=i\hbar\left(x\frac{\partial}{\partial y}-y\frac{\partial}{\partial x}\right)$$

It is a straightforward exercise to check that for the circularly-symmetric Hamiltonian above,

$$[H,L]=0$$

To take advantage of the circular symmetry, we switch to circular variables (r,ϕ) ,where

$$r=\sqrt{x^2+y^2},\ \phi=\tan^{-1}(y/x), so\, x=r\cos\phi, y=r\sin\phi.$$

Transforming the Hamiltonian and angular momentum into (r,ϕ) coordinates,

$$H\psi(r,\phi)=-\frac{\hbar^2}{2M}\left(\frac{\partial^2}{\partial r^2}+\frac{1}{r}\frac{\partial}{\partial r}+\frac{1}{r^2}\frac{\partial^2}{\partial\phi^2}\right)\psi(r,\phi)+V(r)\psi(r,\phi)$$
$$=E\psi(r,\phi)\text{ and }L=-i\hbar\frac{\partial}{\partial\phi}.$$

The angular momentum eigenfunctions $\psi_m(r,\phi)$ satisfy

$$L\psi_m(r,\phi)=-i\hbar\frac{\partial}{\partial\theta}\psi_m(r,\phi)=m\hbar\psi_m(r,\phi)$$

equivalent to $L|m\rangle=m\hbar|m\rangle$. So $\psi_m(r,\phi)=R(r)e^{im\phi}$, and for this to be a single-valued wave function, m must be an integer. (This also ensures the hermiticity of the operator the integration-by-parts check has canceling contributions from $\phi = 0$ and $\phi = 2\pi$.)

Notice this means that *any* function of r multiplied by $e^{im\phi}$ is an eigenfunction of angular momentum with eigenvalue, and in fact any eigenfunction of L with eigenvalue $m\hbar$ must be of this form. So we can factor out the r-dependence, and write a complete set of orthonormal eigenfunctions of L, normalized by integrating around the circle:

$$\Phi_m(\phi) = \frac{1}{\sqrt{2\pi}} e^{im\phi}, \phi \text{ an integer.}$$

It is interesting to note that this would be a complete set of wave functions for a particle confined to a ring rather like the original Bohr orbits. In fact, nanotech rings in which electrons have wave functions like this can now be manufactured. Note also that in such rings one can also have *real* wave functions $\sqrt{1/\pi}\ sin\ m\phi, \sqrt{1/\pi}\ cos\ m\phi$, which are still energy eigenstates, but *not* angular momentum eigenstates, since they are standing waves, linear superpositions of waves going around the ring in opposite directions.

The common eigenstates of the Hamiltonian and the angular momentum evidently have the form

$$|E,m\rangle = \psi_{E,m}(r,\phi) = R_{E,m}(r)\Phi_m(\phi).$$

We should emphasize that although the angular part of the wave function does not depend on the radial potential, the radial component $R_{E,m}(r)$ *does* depend on the angular momentum m. This becomes obvious on putting this $\psi_{E,m}(r,\phi)$ into the (r,ϕ) version of Schrodinger's equation,

$$-\frac{\hbar^2}{2M}\left(\frac{\partial^2}{\partial r^2} + \frac{1}{r}\frac{1}{\partial r} + \frac{1}{r^2}\frac{\partial^2}{\partial \phi^2}\right) R_{E,m}(r)\Phi_m(\phi) + V(r)R_{Em}(r)\Phi_m(\phi)$$

$$= ER_{E,m}(r)\Phi_m(\phi)$$

noting that $\partial^2 / \partial\phi^2 = -m^2$, and canceling out the common factor $\Phi_m(\phi)$ to give

$$-\frac{\hbar^2}{2M}\left(\frac{d^2}{dr^2}+\frac{1}{r}\frac{d}{dr}-\frac{m^2}{r^2}\right)R_{E,m}(r)+V(r)R_{E,m}(r)=ER_{E,m}(r).$$

In this one-dimensional equation for the radial wave function $R_{E,m}(r)$, the angular momentum term $\hbar^2 m^2/2Mr^2 = L^2/2Mr^2$ evidently is equivalent to a repulsive potential. It's called the 'centrifugal barrier' and is easy to understand from classical mechanics. To see this, consider a classical particle bound (in two dimensions) by an attractive central force *V*(*r*). Split the momentum into a radial component p_r and a component in the direction perpendicular to the radius, $p_\perp$. The angular momentum $L = rp_\perp$ and is constant (since the force is central). The energy

$$E=\frac{p^2 r}{2M}+\frac{p^2\perp}{2M}+V(r)=\frac{p^2 r}{2M}+\frac{L^2}{2Mr^2}+V(r)$$

substituting $p_\perp = L/r$. Since $L = m\hbar$, the angular part is exactly equivalent to the above Schrodinger equation.

But what about the radial part? Why isn't *pr*, just equal to $-i\hbar\partial/\partial r$, and $p^2 r$ equal to $\hbar^2\partial^2/\partial r^2$? We know the more complicated differentiation with respect to *r* in the Schrodinger equation above must be correct, because it came from $\partial^2/\partial x^2+\partial^2/\partial y^2$ and $r=\sqrt{x^2+y^2}$, $\phi = tan^{-1}(y/x)$.

To see why p_r equal to $-i\hbar\partial/\partial r$ is incorrect, even though it satisfies $[r, p_r] = i\hbar$, recall what happens in *x*-space. We argued there that $p_x = -i\hbar\partial/\partial x$ for a plane wave because from the photon analogy, acting on the plane wave state $Ce^{ip_x x/k}$ this operator gives the rate of change of phase and therefore the momentum. But a *radial* wave is a little different: picture a photon wave coming out of a single slit having width far less than the photon wavelength. It will radiate outwards with equal amplitude in all directions (180°) but the *wave amplitude will decrease* with distance from the slit to conserve probability. For a long (narrow) slit, this is essentially a two-dimensional problem, so the wave function will be $\psi(r)\cong Ce^{ipr/k}\sqrt{r}$. We know that if we measure the momentum of photons at different distances from the slit we'll get the same result, the wavelength isn't changing, and that determines the phase behaviour. However, the operator $-i\hbar\partial/\partial r$ picks up *an extra term* from differentiating the $\sqrt{r}$, so it is obviously *not* giving us the momentum. Fortunately, this is easy to fix: we define the operator

$$\hat{p}_r = -i\hbar\left(\frac{\partial}{\partial r}+\frac{1}{2r}\right)$$

which eliminates the extra term, *and still satisfies* $[r, pr] = i\hbar$.

However, there is still a small problem. If we substitute this $\hat{p}_r$ in the classical expression for the energy, following the procedure we used successfully to find Schrodinger's equation in Cartesian coordinates, we find

$$H=\frac{p^2r}{2M}+\frac{L^2}{2Mr^2}+V(r)$$
$$=\frac{-\hbar^2\left(\frac{\partial}{\partial r}+\frac{1}{2r}\right)^2}{2M}+\frac{L^2}{2Mr^2}+V(r)$$
$$=\frac{-\hbar^2\left(\frac{\partial^2}{\partial r^2}+\frac{1}{r}\times\frac{\partial}{\partial r}-\frac{1}{4r^2}\right)}{2M}+\frac{L^2}{2Mr^2}+V(r)$$

This is almost but not quite the same as the equation we found by transforming from Cartesian coordinates. The difference is the term $\hbar^2/8Mr^2$. So which is right? Actually our first one was right this second one, derived directly from the classical Hamiltonian, does give the same result in the classical limit, because the difference between them vanishes for $\hbar\to 0$. We conclude that beginning with the classical Hamiltonian, and replacing dynamical variables with the appropriate quantum operators, cannot guarantee that we get the correct quantum Hamiltonian: it might be off by some term of order $\hbar$. This would become evident in predicting properties of truly quantum systems, such as atomic energy levels. Problems of this kind are common in constructing quantum theories starting from a classical theory: essentially, in a classical theory, the order of variables in an expression is irrelevant, but in the quantum theory there can only be *one* correct order of noncommuting variables such as $\partial/\partial r$ and *r* in any expression.

What can we say about the radial wave function $R_{E,m}(r)$? If both the energy and the potential at the origin are finite, then for small r $R_{E,m}(r) \approx Ar^m$ or Ar^{-m}. However, the wave function cannot be discontinuous, so $R_{E,m}(r) \approx Ar^{|m|}$. To make further progress in finding the wave function, we need to know the potential. Specific examples will be analyzed in due course. It is interesting to note that the allowed wave functions, proportional to $r^m e^{im\phi}, {}^{im\phi}, m > 0$, are the complex functions $z^m, \left(z^*\right)^m$ if the two-dimensional space is mapped into the complex plane. Representing many-electron wave functions in the plane in this way was a key to understanding the quantum Hall effect.

Orbital Eigenfunctions in 3-D

The Angular Momentum Operators in Spherical Polar Coordinates

The angular momentum operator $\vec{L} = \vec{r} \times \vec{p} = i\hbar\vec{r} \times \vec{\nabla}$.

In spherical polar coordinates,

$x = r\sin\theta\cos\phi$

$y = r\sin\theta\sin\phi$

$z = r\cos\theta$

$ds^2 = dr^2 + r^2 d\theta^2 + r^2\sin^2\theta d\phi^2$

the gradient operator is

$$\vec{\nabla} = \hat{r}\frac{\partial}{\partial r} + \hat{\theta}\frac{1}{r}\frac{\partial}{\partial\theta} + \hat{\phi}\frac{1}{r\sin\theta}\frac{\partial}{\partial\phi}$$

where now *the little hats denote unit vectors*: $\hat{r}$ is radially outwards, $\hat{\theta}$ points along a line of longitude away from the north pole (and therefore in the direction of increasing θ) and $\hat{\phi}$ points along a line of latitude in an anticlockwise direction as seen looking down on the north pole (that is, in the direction of increasing ϕ). $\hat{r}, \hat{\theta}, \hat{\phi}$ form an orthonormal local basis, and

$$\hat{r} \times \hat{\theta} = \hat{\phi},\ \hat{r} \times \hat{\theta} = -\hat{\theta}$$

so

$$\vec{r} \times \vec{\nabla} = \hat{\phi}\frac{\partial}{\partial\theta} - \hat{\theta}\frac{1}{\sin\theta}\frac{\partial}{\partial\phi}.$$

(Explicitly, $\hat{\phi} = (-\sin\phi, \cos\phi, 0)$ and

$\hat{\theta} = (\cos\theta\cos\phi, \cos\theta\sin\phi, -\sin\theta)$.

The vector $\hat{\phi}$ has zero component in the z-direction, the vector $\hat{\theta}$ has component $-\sin\theta$ in the z-direction, so we can immediately conclude that

$$L_z = (\vec{r} \times \vec{p})_z = -(i\hbar\vec{r} \times \vec{\nabla})_z = i\hbar\frac{\partial}{\partial\phi}$$

just as in the two-dimensional case.

The operator

$$L^2 = -\hbar^2\left(\hat{\phi}\frac{\partial}{\partial\theta} - \hat{\theta}\frac{1}{\sin\theta}\frac{\partial}{\partial\phi}\right)\cdot\left(\hat{\phi}\frac{\partial}{\partial\theta} - \hat{\theta}\frac{1}{\sin\theta}\frac{\partial}{\partial\phi}\right)$$

To evaluate this expression, we use $\hat{\phi}^2 = 1$, $\hat{\theta}^2 = 1$, $\hat{\phi} \times \hat{\theta} = 0$ but we must also check the effects of the differential operators in the first expression on the variables in the second, *including the unit vectors*. Notice $\partial\hat{\phi}/\partial\theta = 0$, $\partial\hat{\theta}/\partial\theta$ is in the r-direction, $\partial\hat{\phi}/\partial\theta$ is a horizontal unit vector pointing inwards perpendicular to $\hat{\phi}$, and having component $-\cos\theta$ in the $\hat{\theta}$-direction, $\partial\hat{\theta}/\partial\phi = \hat{\phi}\cos\theta$. Therefore, the *only* 'differentiation of a unit vector' term that contributes to L^2 is $\hbar^2\hat{\theta}\frac{1}{\sin\theta}\times\frac{\partial\hat{\phi}}{\partial\phi}\frac{\partial}{\partial\theta} = -\hbar^2\cot\theta\frac{\partial}{\partial\theta}$. The $\hat{\phi}\frac{\partial}{\partial\theta}$ acting on the $\sin\theta$ in $-\hat{\theta}\frac{1}{\sin\theta}\frac{\partial}{\partial\phi}$ contributes nothing because $\hat{\phi} \times \hat{\theta} = 0$.

Therefore

$$L^2 = -\hbar^2\left(\frac{\partial^2}{\partial\theta^2} + \cot\theta\frac{\partial}{\partial\theta} + \frac{1}{\sin^2\theta}\frac{\partial^2}{\partial\phi^2}\right)$$

$$= -\hbar^2\left(\frac{1}{\sin\theta}\frac{\partial}{\partial\theta}\sin\theta\frac{\partial}{\partial\theta} + \frac{1}{\sin^2\theta}\frac{\partial^2}{\partial\phi^2}\right)$$

Now, we know that L^2 and L_z have a common set of eigenkets (since they commute) and we've already established that those of L_z

are $\Phi(\phi) = e^{im\phi} / \sqrt{2\pi}$, with m an integer, so the eigenkets of L^2 must have this same ϕ dependence, so they must be of the form $\Theta_l^m(\theta)\Phi(\phi)$, where $\Theta_l^m(\theta)$ is a (suitably normalized) solution of the equation

$$\frac{1}{\sin\theta}\frac{d}{d\theta}\sin\theta\frac{d\Theta_l^m(\theta)}{d\theta} - \frac{m^2}{\sin^2\theta}\Theta_l^m(\theta) = -l(l+1)\Theta_l^m(\theta)$$

more conveniently written

$$\sin\theta\frac{d}{d\theta}\sin\theta\frac{d\Theta_l^m(\theta)}{d\theta}\left(l(l+1)\sin^2\theta - m^2\right)\Theta_l^m(\theta) = 0.$$

Finding the m = l Eigenket of L^2, L_z

The easiest wave function to find was that of the ground state, the solution of the linear equation $\hat{a}\psi_0 = 0$, and the other state wave functions could then be found by applying the creation operator in differential form the necessary number of times.

A similar strategy works here: we find the *highest* state on the l ladder, $m = l$, the state $|l,l\rangle$, from the equation $L_+|l,l\rangle = 0$, where $L_+ = L_x + iL_y$. So we need to find $L_+ = i\hbar\left(\vec{r}\times\vec{\nabla}\right)_+$.

From $\vec{r}\times\vec{\nabla} = \hat{\phi}\frac{\partial}{\partial\theta} - \hat{\theta}\frac{1}{\sin\theta}\frac{\partial}{\partial\phi}$,

we have

$$(\vec{r}\times\vec{\nabla})_+ = \hat{\phi}\frac{\partial}{\partial\theta} - \hat{\theta}_+\frac{1}{\sin\theta}\frac{\partial}{\partial\phi},$$

and using

$\hat{\phi} = (-\sin\phi, \cos\phi, 0), \hat{\theta} = (\cos\theta\cos\phi, \cos\theta\sin\phi, -\sin\theta)$, we see that $\hat{\phi}_+$, the component of $\hat{\phi}$ in the + direction, is $\hat{\phi}_+ = \hat{\phi}_x + \hat{\phi}_y = ie^{i\phi}$, and similarly $\hat{\theta}_+ = \cos\theta e^{i\phi}$.

Therefore,

$$L_+ = \hbar e_{i\phi}\left(\frac{\partial}{\partial\theta} + i\cot\theta\frac{\partial}{\partial\phi}\right)$$

$$L_+ = \hbar e_{i\phi}\left(\frac{\partial}{\partial\theta} - i\cot\theta\frac{\partial}{\partial\phi}\right)$$

So, $L_+ \left| l,l \right\rangle = 0$ becomes

$$\left(\frac{\partial}{\partial\theta} + i\cot\theta\frac{\partial}{\partial\phi} \right)\Theta_l^l(\theta)e^{il\phi}$$

that is,

$$\left(\frac{d}{d\theta} - l\cot\theta \right)\Theta_l^l(\theta) = 0.$$

The solution to this equation is

$$\Theta_l^l(\theta) = N(\sin\theta)^l \quad .$$

where N is the normalization constant. The $m \neq l$ wave functions are generated by applying the lowering operator L_-

Normalizing the m = l Eigenket

The standard notation for the normalized eigenkets $\left| l,m \right\rangle$ is $Y_l^m(\theta,\phi) = \Theta_l^m(\theta)\Theta_m(\phi)$. These functions, being eigenkets of Hermitian operators with different eigenvalues, must satisfy

$$\int_{\theta-0}^{x}\int_{\phi-0}^{2x}(\theta,\phi)Y_l^m(\theta,\phi)\sin\theta d\theta d\phi = \int Y_l^{m'*}(\theta,\phi)Y_l^m(\theta,\phi)d\Omega = \delta_{ll}\delta_{mm}.$$

So, our first job is to normalize $\Theta_l^l(\theta) = N(\sin\theta)^l$ (taking $\Phi_l(\phi) = e^{il\phi}/\sqrt{2\pi}$ already normalized)

$$\left| N^2 \right| \int_0^z (\sin\theta)^{2l+1} d\theta = 1$$

The integral can be evaluated using the substitution $\mu = \cos\theta$ to give $\int_{-1}^{1}\left(1-\mu^2\right)^l d\mu$, then making the further substitution $u = \frac{1}{2}(1-\mu)$ to give $2^{2L+1}\int_0^1 u^1(1-u)^l du$, which can be integrated by parts to give

$$|N|^2 2^{2l+1}(l!)^2(2l+1)! = 1 .$$

Therefore,

$$Y_l^l(\theta,\phi) = (-1)^l\left(\frac{(2l+1)!}{4\pi}\right)\frac{1}{2^l l!}(\sin\theta)^l e^{il\phi} = c_l(\sin\theta)^l e^{il\theta}$$

where we have fixed the sign in accord with the standard convention, and we will denote the rather cumbersome normalization constant by c_l.

Notice that for large values of l, this function is heavily weighted around the equator, as we would expect for a given total angular momentum one gets a maximum component in the z-direction when the motion is concentrated in the x, y plane. This looks like a Bohr orbit.

Finding the Rest of the Eigenkets

Now that $|l,l\rangle$ is normalized, we can automatically produce correctly normalized $|l,m\rangle$'s, since we know the matrix element of the lowering operator between normalized states. We don't have to do any more integrals.

For example, $L_-|l,l\rangle = \hbar\sqrt{2l}\,|l,l-1\rangle$, equivalently (the $\hbar$'s of course cancel)

$$Y_l^{l-1}(\theta-\phi)=\frac{(-1)}{\sqrt{2l}}e^{-i\phi}\left(\frac{\partial}{\partial\theta}-i\cot\theta\frac{\partial}{\partial\phi}\right)Y_l^l$$

That is,

$$Y_l^{l-1}(\theta-\phi)=c_l e^{-i\phi}\left(\frac{\partial}{\partial\theta}-i\cot\theta\frac{\partial}{\partial\phi}\right)\sin^l\theta\times e^{il\phi}$$

$$=c_l e^{-i(l-1)\phi}\sqrt{2l}\sin^{l-1}\theta\cos\theta$$

(both terms giving equal contributions).

Note that this function is actually *zero* on the equator, but for large l it peaks close to the equator (on both sides).

In principle, we can reapply this differential operator over and over to generate all the $|l,m\rangle$ states, but this gets very messy. However, there is a neat theorem concerning the lowering operator that makes it all straightforward:

$$L_- e^{im\phi}f(\theta)=e^{i(m-1)\phi}\left(\sin^{1-m}\theta\frac{d}{d(\cos\theta)}\sin^m\theta\right)f(\theta)$$

$$\text{So, } L_- e^{im\phi}\sin^l(\theta) = e^{i(l-1)\phi}\left(\sin^{1-l}\theta\frac{d}{d(\cos\theta)}\sin^l\theta\right)\sin^l(\theta)$$

and applying the operator again,

$$\left(L_-^{\ 2}\right)e^{im\phi}\sin^l(\theta) = L_- e^{i(l-1)\phi}\left(\sin^{1-l}\theta\frac{d}{d(\cos\theta)}\sin^l\theta\right)\sin^l(\theta)$$

$$= e^{i(l-2)\phi}\left(\sin^{2-l}\theta\frac{d}{d(\cos\theta)}\sin^{l-1}\theta\right)\left(\sin^{1-l}\theta\frac{d}{d(\cos\theta)}\sin^l\theta\right)\sin^l(\theta)$$

$$= e^{i(l-2)\phi}\left(\sin^{2-l}\theta\frac{d^2}{d^2(\cos\theta)}\sin^l\theta\right)\sin^l(\theta).$$

So the point of introducing this odd-looking representation of the lowering operator is that the $\sin^{l-l}$ term in the middle is exactly *cancelled* when the operator is applies twice, and similar cancellations occur on repeating the operation, giving the (relatively) simple representation:

$$Y_l^m(\theta,\phi) = c_l\sqrt{\frac{(1+m)!}{(2l)!(1-m)!}}e^{-im\phi}\sin^{-m}\theta\frac{d^{l-m}}{d(\cos\theta)^{l-m}}\sin^{2l}\theta$$

(Where did all those factorials come from? They're the product of all the inverse square root factors in

$|l,m-1\rangle = \frac{1}{\sqrt{(1+m)(1-m+1)}}L_-|l,m\rangle$ for the number of lowerings necessary.)

Note that for $m = 0$ the function is

$$Y_l^0(\theta,\phi) = c_l\sqrt{\frac{1}{(2l)!}}\frac{d^l}{d(\cos\theta)^l}\sin^{2l}\theta,$$

and in fact not a function of ϕ at all. This isn't surprising, since it has zero angular momentum about the z-direction, the appropriate $\Phi(\phi)$ is just constant.

For m=-$l$the differentiation becomes trivial, because, writing $\cos\theta = \mu$, the differentiation becomes $\frac{d^{2l}}{d\mu^{2l}}\left(1-\mu^2\right)^l$ and only the term survives, giving

$$Y_l^{-l}(\theta,\phi)=(-1)^l c_l e^{-il\phi}\sin^l\theta.$$

Of course, this could also have been found from the linear equation $L_-|l,-l\rangle=0$, and we could have equally generated all the states by applying L_+ to this state. In fact, this gives a different but of course equivalent expression for the $y_l^m(\theta,\phi)$:

$$y_l^m(\theta,\phi)=(-1)ct\sqrt{\frac{(1-m)!}{(2l)!(1+m)!}}e^{im\phi}\sin^m\theta\frac{d^{l+m}}{d(\cos\theta)^{l+m}}\sin^{2l}\theta$$

Relating the $Y_l{}^m$'s to the Legendre Functions

The *Legendre polynomials* $P_n(\cos\theta)$ are defined by:

$$P_n(\cos\theta)=\frac{1}{2^n n!}\frac{d^n}{d(\cos\theta)^n}\sin^{2n}\theta, \text{ or}$$

$$P_n(\mu)=\frac{1}{2^n n!}\frac{d^n}{d\mu^n}(1-\mu^2)^2.$$

where $\mu=\cos\theta$, so $d\mu=-\sin\theta d\theta$. From this form, it is easy to show that $P_n(1)=1$ (all n differentiations must take out a $(1-\mu^2)$ factor to give a nonzero contribution), and $p_n(\mu)$ must have n zeros in the interval (-1, 1). $p_n(\mu)$ alternates between an even function and an odd function.

The normalization of the $p_n(\mu)$'s is

$$=\int_{-1}^{1}(P_n(\mu))^2 d\mu=\left(\frac{1}{2^n n!}\right)^2\int_{-1}^{1}\frac{d^2}{d\mu^n}(\mu^2-1)^2 d\mu$$

$$=(-1)\left(\frac{1}{2^n n!}\right)^2\int_{-1}^{1}\frac{d^2}{d\mu^n}(\mu^2-1)^n\frac{d^n}{d\mu^n}(\mu^2-1)^2 d\mu$$

$$=(2n)!\left(\frac{1}{2^n n!}\right)^2\int_{-1}^{1}(\mu^2-1)^n d\mu$$

$$=\frac{2}{2n+1}$$

where in that last line we used the result for the integral obtained earlier in this lecture for normalizing Y_l^l.

Doing the same repeated integration by parts for two different Legendre polynomials proves they are orthogonal,

$$\int_{-1}^{1} P_m(\mu)P_n(\mu)d\mu = 0, \quad m \neq n.$$

The *associated Legendre functions* are defined (for n and m zero or positive integers, $n \geq m$) by:

$$P_n^m(\mu) = (1-\mu)^{m/2}\frac{d^m}{d\mu^m}P_m(\mu)$$

$$= (-1)^n \frac{\left(1-\mu^2\right)^{m/2}}{2^n n!}\frac{d^{n+m}}{d\mu^{n+m}}\left(1-\mu^2\right)^n.$$

Following Messiah in requiring $Y_l^0(0,0)$ be real and positive, we find

$$Y_l^0(0,\theta) = \sqrt{\frac{2l+1}{4\pi}}\, p_l(\cos\theta)$$

where the coefficient just reflects the differing normalization conventions. Similarly, the spherical harmonics with nonzero m are proportional to the associated Legendre functions (the odd ones are *not* polynomials in $\cos\theta$.

The Spherical Harmonics as a Basis

We have found explicit expressions for the spherical harmonics: an orthonormal set of eigenfunctions of L^2 and L_z defined on the surface of a sphere,

$$\int_{\theta-0}^{:}\int_{\phi-0}^{2:} Y_l^{m'*}(\theta,\phi)Y_l^m(\theta,\phi)\sin\theta d\theta d\phi = \int Y_l^{m'*}(\theta,\phi)Y_l^m(\theta,\phi)d\Omega$$

$$\delta_{ll}\delta_{mm'}.$$

They form a *complete* set:

$$\sum_{l=0}^{\infty}\sum_{m-1}^{l}|l,m\rangle\langle l,m| = 1$$

or

$$\sum_{l=0}^{\infty}\sum_{m-1}^{l} Y_l^{m*}(\theta,\phi)Y_l^m(\theta',\phi') = \delta(\cos\theta-\cos\theta')\delta(\phi-\phi') = \delta(\Omega-\Omega')$$

in the notation of Messiah, where Ω refers to a point on the spherical surface.

(Formal proof of the completeness is given in Byron and Fuller, *Mathematics of Classical and Quantum Physics*.)

The above equation could also be written

$$\sum_{l-0}^{\infty}\sum_{m--l}^{l}\langle\theta,\phi|1,m\rangle\langle 1,m|\theta',\phi\rangle=\langle\theta,\phi|\theta',\phi'\rangle$$

$$\delta(\cos\theta-\cos\theta')\delta(\phi-\phi')$$

where the ket $|\theta',\phi'\rangle$ is to be understood as a localized ket, the spherical-surface version of $|x\rangle$, normalized by its δ-function inner product with the bra $|\theta,\phi\rangle$, exactly analogous to $\langle x|x'\rangle=\delta(x-x')$, bearing in mind that the infinitesimal area element is $-d(\cos\theta)d\phi$, (a positive quantity in the relevant interval, 0 to π).

This completeness means that any reasonable function on the surface of the sphere can be expressed as a sum over spherical harmonics with appropriate coefficients, in other words the spherical generalization of a Fourier series.

In fact, L^2 is equivalent to ∇^2 on the spherical surface, so the Y_l^m are the eigenfunctions of the operator ∇^2. Just as in one dimension the eigenfunctions of d^2/dx^2 have the spatial dependence of the eigenmodes of a vibrating string, the spherical harmonics have the spatial dependence of the eigenmodes of a vibrating spherical balloon. Of course, to describe the displacement of the balloon skin (which must be real!) with these eigenfunctions, we can no longer use the eigenfunctions of the z-component of angular momentum, since they are complex except in the trivial zero case. We must rearrange the eigenfunctions of L^2, for example replacing the pair $e^{i\phi},e^{-i\phi}$ with $\cos^{\phi},\sin^{\phi}$. These real solutions, essentially $\frac{1}{\sqrt{2}}(|l,l\rangle\pm|l,-l\rangle)$, have l nodal lines (zeroes) of longitude. Moving down one notch in $|m|$, the (real) state with $|m|=l-1$ has l-1 longitudinal nodes, but has added a *latitudinal* node: the equator. Then $|m|=l-2$ has l-2 longitudinal nodes, 2 latitudinal nodal lines *there are always l nodal lines total.*

Some of these modes of vibration have been observed in the sun after a sunspot storm. The spherical harmonics are also used in analyzing the cosmic background radiation.

Some Low Order Spherical Harmonics

Let's look in more detail at the lowest order spherical harmonics, from our general formulas.

$$Y_0^0 = \frac{1}{\sqrt{4\pi}}$$

$$Y_1^1 = \sqrt{\frac{3}{8\pi}} \sin\theta e^{i\phi}$$

$$Y_1^0 = \sqrt{\frac{3}{4\pi}} \cos\theta$$

$$Y_1^{-1} = \sqrt{\frac{3}{8\pi}} \sin\theta e^{-i\phi}$$

$$Y_2^2 = \sqrt{\frac{15}{32\pi}} \sin^2\theta e^{2i\phi},\ Y_2^1 = -\sqrt{\frac{15}{8\pi}} \sin\theta \cos\theta e^{i\phi},$$

$$Y_2^0 = \sqrt{\frac{5}{16\pi}} 3\left(\cos^2\theta - 1\right)$$

$$Y_2^{-2} = \sqrt{\frac{15}{32\pi}} \sin^2\theta e^{-2i\phi}, Y_2^{-1} = -\sqrt{\frac{15}{8\pi}} \sin\theta \cos\theta e^{i\phi}$$

It is often useful to write the Y_l^m in terms of Cartesian coordinates,

$$(x, y, z) = (r \sin\theta \cos\phi, r \sin\theta \sin\phi, r \cos\theta)$$

so

$$Y_1^1(x, y, z) = -\sqrt{\frac{3}{8\pi}} \times \frac{x + iy}{r}, Y_1^0(x, y, z) = \sqrt{\frac{3}{4\pi}} \times \frac{z}{r}$$

$$(x, y, z) = \sqrt{\frac{3}{8\pi}} \times \frac{x - iy}{r}$$

and

$$Y_2^2 = \sqrt{\frac{15}{32\pi}} (x + iy)^2, Y_2^1 = -\sqrt{\frac{15}{8\pi}} (x + iy) z, Y_2^0 = \sqrt{\frac{5}{16\pi}} \left(3z^2 - 1\right), etc.$$

The Y_1^m as a Basis of the $l = 1$ Subspace

The Y_1^m are the $l = 1$ eigenstates of L^2 and L_z. But what if we'd chosen to look for the common eigenstates of L^2 and L_x instead? What $l = 1$ state has zero angular momentum component in the direction of the x-axis? Clearly it will be $\sqrt{\frac{3}{4\pi}} \times \frac{x}{r}$, in other words the previous Y_1^0 with z replaced by x, because after all, our labeling of axes was arbitrary.

$$\text{Now, } \sqrt{\frac{3}{4\pi}} \times \frac{x}{r} \quad \text{is just} \left(1+\sqrt{2}\right)\left(-Y_1^1 + Y_1^{-1}\right)$$

In fact, *any* $l = 1$ state, with a specified component in *any* direction, can be written as

$$\alpha_1 |1,1\rangle + \alpha_0 |1,0\rangle + \alpha_{-1} |1,-1\rangle = \sum \alpha_m |1,m\rangle.$$

This can be seen as follows: an $l = 1$ state has to be *linear* in $x/r, y/r, z/r$ (any quadratic term would give rise to about an appropriate axis, call that the z-axis, so $m = 2$ and l must be 2 or greater), and any such state can be written as a linear combination of $(x+iy)/r, (x-iy)/r, z/r$.

The bottom line, then, is that the Y_1^m do indeed provide a *complete* basis for the $l = 1$ space of eigenstates of L^2.

Representing the Rotation Operator within the $l = 1$ Subspace

Recall that we originally introduced the angular momentum operator by defining it as the generator of infinitesimal rotations when acting on any wave function, including multicomponent wave functions. We found, using the commutativity properties of ordinary rotations, that the vector components $\vec{J}$ of had to satisfy $\left[J_x, J_y\right] = i\hbar J_z$, etc., and from that we deduced the possible sets of eigenvalues of the commuting pair of operators $\vec{J}^2, J_x$ were $j(j+1)\hbar^2$ for $\vec{J}^2$, with j an integer of half an odd integer, and for each such j the allowed eigenvalues of J_z were $m\hbar, m = -j, -j+1, \ldots, +j$.

Back to the $l = 1$ angular wave functions: we have established that any such function can be written $\alpha_1|1,1\rangle + \alpha_0|1,0\rangle + \alpha_{-1}|1,-1\rangle = \sum \alpha_m|1,m\rangle$, and so is a vector in a three-dimensional space spanned by the set $|l,m\rangle$. In other words, the wave function is a three-component object. The angular momentum operator must therefore be *a matrix operator in this three-dimensional space*, such that, by definition, the effect of an infinitesimal rotation on the multicomponent wave function is:

$$R(\delta\vec{\theta})\psi_{1\text{-}1}(\theta,\phi) = e^{-\frac{1}{k}\delta\vec{\theta}\hat{J}}\begin{pmatrix}\alpha_1\\ \alpha_0\\ \alpha_{-1}\end{pmatrix} = \begin{pmatrix}\alpha_1'\\ \alpha_0'\\ \alpha_{-1}'\end{pmatrix}$$

The unitary rotation operator acting in the $l = 1$ subspace, $U(R(\vec{\theta})) = e^{-\frac{i\vec{\theta}.\vec{J}}{k}}$, has to be a 3 × 3 matrix. The standard notation

for its matrix elements is:

$$D^{(1)}_{m'm}(R(\vec{\theta})) = \langle 1,m'|e^{-\frac{i\vec{\theta}\,\vec{J}}{k}}|1,m\rangle$$

so the rotated ket is

$$\alpha_{m'}' = \sum_{m',m} D^{(1)}_{m'm}\alpha_m,\ or\,\alpha' = D\alpha.$$

To evaluate this matrix explicitly, we must expand the exponential and we need the matrix elements of J_z, J_+, J_- between the states $|l,m\rangle$ which we already know.

Now, the basis of the three-dimensional space is just the common eigenkets of $\vec{J}^2, J_z$, in this case identical to $\vec{J}^2, J_z$. We know the matrix elements of J_z, J_+, J_- between states from the earlier lecture, so it is simple to find the matrix $|J,m\rangle$ representations of the components of J in this space:

$$J_x^{(1)} = \frac{\hbar}{\sqrt{2}}\begin{pmatrix}0 & 1 & 0\\ 1 & 0 & 1\\ 0 & 1 & 0\end{pmatrix},\ J_y^{(1)} = \frac{i\hbar}{\sqrt{2}}\begin{pmatrix}0 & -1 & 0\\ 1 & 0 & -1\\ 0 & 1 & 0\end{pmatrix}$$

$$J_z^{(1)} = \hbar\begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -1 \end{pmatrix}$$

We have added the superscript (1) because this representation of the infinitesimal rotation operators is specific to $j = 1$ (representations for general values of j are as $(2j + 1) \times (2j + 1)$matrices, reflecting the dimensionality of the space spanned by the $2j + 1$ distinct m values).

Expanding the exponential is not difficult, because by inspection $\left(J_z^{(1)}/\hbar\right)^3 = \left(J_z^{(1)}\hbar\right)$, so from spherical symmetry $\left(\hat{\vec{n}} \times \vec{J}^{(1)}/\hbar\right)^3 = \left(\hat{\vec{n}} \times \vec{J}^{(1)}/\hbar\right)$ for a unit vector in any direction. The result is:

$$D^{(1)}\left(R\left(\vec{\theta}\right)\right) = e^{-\frac{i\theta\hat{\vec{n}}\vec{J}}{k}} = 1 + \left(\cos\theta - 1\right)\left(\frac{\hat{\vec{n}} \times \vec{J}}{\hbar}\right)^2 - i\sin\theta\left(\frac{\hat{\vec{n}} \times \vec{J}}{\hbar}\right).$$

6

Niels Bohr and Quantum Atom

So far we have seen that quantum physics deals with electromagnetic radiation — that is to say, light. But at the beginning we said that quantum physics tells us that *material things* are described by quantum physics. So what happened to matter?

Niels Bohr happened to matter. Bohr was a Danish physicist whose parents were both scientists. Apparently young Niels grew up in an atmosphere that was favourable to science. He received his doctorate in 1911 from the University of Copenhagen.

One can see just by looking at Niels Bohr's early academic life (though there are many other indicators) that the 'atomic theory' of matter — that is to say that matter is made up of tiny atoms — had gone from being scientifically marginalised to being at the heart of physics in a very short time. Bohr's doctoral thesis was concerned with the inadequacy of classical (that is to say, Newtonian) physics for describing the behaviour of matter at the atomic level.

In the 19th century, physicists who saw some value in the atomic theory thought of the atom as a tiny undivided and indivisible unit of matter, the smallest possible unit into which something could be broken down. Experiments dealing with the photoelectric effect, along with other observations, strongly suggested that the atom had some internal structure, since particles called 'electrons' were being emitted from them.

Rutherford's Model and Its Drawbacks

A model of the atom was described by the British physicist Ernest Rutherford in 1911, and is known as the Solar System model.

It is very simple, and is still used to teach elementary atomic structure to school children.

1. An atom consists of a *central nucleus*. This nucleus is composed of *positively charged protons*, and electrically uncharged (*neutral*) *neutrons*.
2. *Negatively charged electrons revolve round the nucleus in definite orbits.*
3. The orbits themselves can be *at any distance* from the nucleus.
4. In any atom, *the number of protons is equal to the number of electrons*, and hence it is electrically neutral.

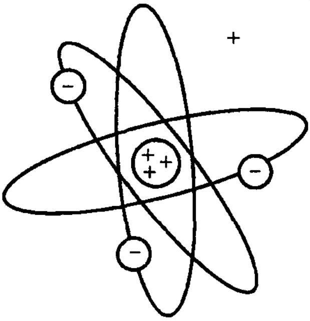


Fig. 1 Rutherford's Atomic Model.

Rutherford used the laws of motion that had been set forward by Sir Isaac Newton to describe the atom. According to Rutherford's description, the electrons of an atom could occupy one of an infinite number of orbits, in accordance with Newton's laws. There were problems with Rutherford's description of the atom from the beginning. Let us find out two drawbacks of Rutherford's theory.

Inherent Instability of the Atom

According to Rutherford's theory, electrons could orbit the nucleus at any distance. When the electrons circle round the nucleus, they are constantly changing their direction. According to classical electrodynamics (which deals with the motion of electrons), such electrons which either constantly change their direction or their velocity or both should continuously emit radiation. While doing so, they should lose energy, and thus spiral into the nucleus. This means *every atom is unstable*, quite contrary to our observation.

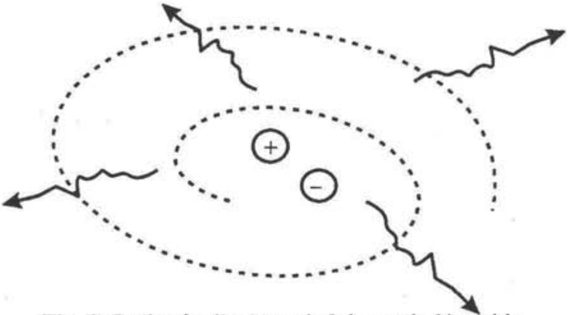


Fig. 2 Rutherford's Atom is Inherently Unstable.

Atomic Spectra

Rutherford's description of the atom could not be entirely correct because it did not account for some observations that had already been made. Perhaps the most important of these observations concerned the behaviour of certain gases. These gases at low pressure emit light in a set of *discrete bands* of the electromagnetic spectrum. This is quite different from the radiation emitted by solids, which is spread evenly across the electromagnetic spectrum. The radiation emissions of these gases were important because they showed that at least under some circumstances, the orbits of the electrons could not be at just any distance from the nucleus, but were confined to discrete distances (or *energy states*).

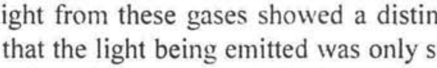


Fig. 3 Continuous Spectrum and B. line Spectrum of Hydrogen.

If the electrons in these gases were free to orbit at any distance, then the light emitted from them would have been spread evenly across the electromagnetic spectrum. Instead, what experimenters saw was that the light from these gases showed a distinct *line pattern.* That is to say that the light being emitted was only seen in a certain set of wavelengths, with empty spaces in between.

These line-spectra were different for each gas, and was found to be the characteristic of its atom. Today, astronomers use line-spectra to detect the elements present in stars.

Bohr's Explanation

Niels Bohr quickly seized upon this problem and used it to propose a *quantized* description of the atom.

1. Bohr proposed that while circling the nucleus of the atom, electrons could only occupy certain *discrete* orbits, that is to say, energy levels. Bohr used Max Planck's equations describing quanta of radiation to determine what these discrete orbits would have to be. *As long as electrons stay in these energy levels, they are stable.*
2. Further, Bohr said electrons give or take energy only when they change their energy levels. If they move up, they *take* energy (say from light), and if they move down, they *release* energy.
3. Furthermore, Bohr also said that an electron which is not in its native energy level (in other words, which has been *excited* to a higher energy level) *always* has to fall back to its original, stable level.

Bohr interpreted the lines in the spectra of gases as formed by the *transitions* of electrons to and from various energy-levels. This has been verified thoroughly with the hydrogen atom, *and found to be correct*. Bohr's formulae agreed excellently with observed line positions.

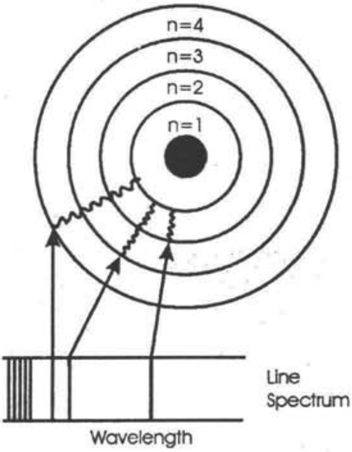


Fig. 4 Bohr's Explanation of Line Spectra.

Imagine that you are taking a walking along a beach. As you walk along, you see a sand-castle that someone has built. As you get closer to the sand castle, you discover that you can only stand three metres, two metres, or one metre from the sand castle. You cannot stand at one and a half meters, nor can you stand at two and three quarters of a metre from the sand castle. No matter how hard you try, some mysterious force keeps you at one of those three distances. In everyday life such a situation is absurdly impossible. In the physics of the very small, it is a necessity.

This description in which electrons can only occupy certain orbits is called the *shell model* of the atom, because Bohr described the possible orbits of the electrons as *orbitals* or *shells*. When an atom of a gas *released* energy, an electron would move *down* to a lower orbit (requiring less energy), and when an atom *acquired* energy, an electron would move *up* to a higher energy level. But these orbits or shells were *discrete*, like the distances from the sand castle. The orbits were not a smooth, continuous series of possibilities as one finds in the everyday world, but rather a set of *distinct states* separated from each other, much like the separation of the quanta of electromagnetic radiation that Planck had discovered. This caused the *distinct lines* in the spectrum.

The Simple Harmonic Oscillator

The simple harmonic oscillator, a nonrelativistic particle in a potential $\frac{1}{2}kx^2$ is an excellent model for a wide range of systems in nature. In fact, not long after Planck's discovery that the black body radiation spectrum could be explained by assuming energy to be exchanged in quanta, Einstein applied the same principle to the simple harmonic oscillator, thereby solving a long-standing puzzle in solid state physics the mysterious *drop in specific heat* of all solids at low temperatures. Classical thermodynamics, a very successful theory in many ways, predicted no such drop with the standard equipartition of energy, *kT* in each mode (potential plus kinetic), the specific heat should remain more or less constant as the temperature was lowered (assuming no phase change).

To explain the anomalous low temperature behaviour, Einstein assumed each atom to be an independent (quantum) simple harmonic oscillator, and, just as for black body radiation, he assumed the oscillators could only absorb or emit energy in *quanta*. Consequently, at low enough temperatures there is rarely sufficient energy in the ambient thermal excitations to excite the oscillators, and they freeze out, just as blue oscillators do in low temperature black body radiation. Einstein's picture was later somewhat refined the basic set of oscillators was taken to be standing sound wave oscillations in the solid rather than individual atoms (making the picture even more like black body radiation in a cavity) but the main conclusion the drop off in specific heat at low temperatures was not affected.

The Classical Simple Harmonic Oscillator

The classical equation of motion for a one-dimensional simple harmonic oscillator with a particle of mass m attached to a spring having spring constant k is

$$m\frac{d^2x}{dt^2}=-kx.$$

The solution is

$$x=x_0\sin\left(\omega t+\delta\right),\ \omega=\sqrt{\frac{k}{m}},$$

and the momentum $p = mv$ has time dependence

$$p=mx_0\omega\cos\left(\omega t+\delta\right).$$

The total energy

$$\left(1/2m\right)\left(p^2+m^2\omega^2x^2\right)=E$$

is clearly constant in time.

It is often useful to picture the time-development of a system in *phase space*, in this case a two-dimensional plot with position on the x-axis, momentum on the y-axis. Actually, to have (x,y) coordinates with the same dimensions, we use $\left(m\omega x, p\right)$.

It is evident from the above expression for the total energy that in these variables the point representing the system in phase space moves clockwise around a circle of radius $\sqrt{2mE}$ centered at the origin.

Note that in the *classical* problem we could choose any point $\left(m\omega x, p\right)$, place the system there and it would then move in a circle about the origin. In the *quantum* problem, on the other hand, we cannot specify the initial coordinates $\left(m\omega x, p\right)$ precisely, because of the uncertainly principle. The best we can do is to place the system initially in a small cell in phase space, of size $\Delta x \times \Delta p = \hbar/2$. In fact, we shall find that in quantum mechanics phase space is always divided into cells of essentially this size for each pair of variables.

SchrodingJer's Equation and the Ground State Wave Function

From the classical expression for total energy given above, the Schrodinger equation for the quantum oscillator follows in standard fashion:

$$-\frac{\hbar^2}{2m}\frac{d^2\psi(x)}{dx^2}+\frac{1}{2}m\omega^2x^2\psi(x)=E\psi(x)$$

What will the solutions to this Schrodinger equation look like? Since the potential $\frac{1}{2}m\omega^2x^2$ increases without limit on going away from $x = 0$, it follows that no matter how much kinetic energy the particle has, for sufficiently large x the potential energy dominates, and the (bound state) wavefunction decays with increasing rapidity for further increase in x. (Obviously, for a real physical oscillator there is a limit on the height of the potential we will assume that limit is much greater than the energies of interest in our problem.)

We know that when a particle penetrates a barrier of constant height V_0 (greater than the particle's kinetic energy) the wave function decreases exponentially into the barrier, as $e^{-\alpha x}$, where $\alpha = \sqrt{2m(V_0 - E)/\hbar^2}$. But, in contrast to this constant height barrier, the 'height' of the simple harmonic oscillator potential *continues to increase* as the particle penetrates to larger x. Obviously, in this situation the decay will be faster than exponential. If we assume it is more or less *locally* exponential, but with a local α varying with V_0, neglecting E relative to V_0 in the expression for α suggests that α itself is proportional to x (since the potential is proportional to x^2, and $\alpha \propto \sqrt{V}$) so maybe the wavefunction decays as $e^{-(constant)x^2}$?

To check this idea, we insert $\psi(x) = e^{-x^2/2b^2}$ in the Schrodinger equation, using

$$\frac{d^2\psi}{dx^2}=-\frac{1}{b^2}\psi+\frac{x^2}{b^2}\psi$$

to find

$$-\frac{\hbar^2}{2m}\left(-\frac{1}{b^2}+\frac{x^2}{b^4}\right)\psi(x)+\frac{1}{2}m\omega^2x^2\psi(x)=E\psi(x).$$

The $\psi(x)$ is just a factor here, and it is never zero, so can be cancelled out. This leaves a quadratic expression which must have the same coefficients of x^0, x^2 on the two sides, that is, the coefficient of x^2 on the left hand side must be zero:

$$\frac{\hbar^2}{2mb^4}=\frac{m\omega^2}{2}, so\, b=\sqrt{\frac{\hbar}{m\omega}}.$$

This fixes the wave function. Equating the constant terms fixes the energy:

$$E=\frac{\hbar^2}{2mb^2}=\frac{1}{2}\hbar\omega.$$

So the conjectured form for the wave function is in fact the *exact* solution for the lowest energy state! (It's the lowest state because it has no nodes.)

Also note that even in this ground state the energy is *nonzero*, just as it was for the square well. The central part of the wave function must have some curvature to join together the decreasing wave function on the left to that on the right. This 'zero point energy' is sufficient in one physical case to melt the lattice helium is liquid even down to absolute zero temperature (checked down to microkelvins!) because the wave function spread destabilizes the solid lattice that will form with sufficient external pressure.

Higher Energy States

It is clear from the above discussion of the ground state that $b=\sqrt{\frac{\hbar}{m\omega}}$ is the natural unit of length in this problem, and that of energy, so to investigate higher energy states we reformulate in dimensionless variables,

$$\xi=\frac{x}{b}=x\sqrt{\frac{m\omega}{\hbar}}, \quad \omega=\frac{E}{\hbar\omega}.$$

Schrodinger's equation becomes

$$\frac{d^2\psi(\xi)}{d\xi^2}=\left(\xi^2-2\varepsilon\right)\psi(\xi).$$

Deep in the barrier, the ε term will become negligible, and just as for the ground state wave function, higher bound state wave functions will have $e^{-\varsigma^2/2}$ behaviour, multiplied by some more slowly varying factor (it turns out to be a polynomial).

Exercise: find the relative contributions to the second derivative from the two terms in $x^n e^{-x^2/2}$ For given n, when do the contributions involving the first term become small? Define 'small'. The standard approach to solving the general problem is to factor out the $e^{-\varsigma^2/2}$ term, $\psi(\xi) = h(\xi)e^{-\varsigma^2/2}$ giving a differential equation for $h(\xi)$:

$$\frac{d^2h}{d\xi^2} - 2\xi\frac{dh}{d\xi} + (2\omega - 1)h = 0.$$

We try solving this with a power series in

$$\xi : h(\xi) = h_0 + h_1\xi + h_2\xi^2 = ...$$

Inserting this in the differential equation, and requiring that the coefficient of each power ξ^n vanish identically, leads to a recurrence formula for the coefficients h_n:

$$h_{n+2} = \frac{(2n+1-2\omega)}{(n+1)(n+1)}h_n.$$

Evidently, the series of odd powers and that of even powers are independent solutions to Schrodinger's equation. (Actually this isn't surprising: the potential is even in x, so the parity operator P commutes with the Hamiltonian. Therefore, unless states are degenerate in energy, the wave functions will be even or odd in x.) For large n, the recurrence relation simplifies to

$$h_{n+2} \approx \frac{2}{n}h_n, \quad n \gg \varepsilon.$$

The series therefore tends to

$$\sum \frac{2^n \xi^{2n}}{(2n-2)(2n-4)...4} = 2\xi^2 \sum \frac{\xi^{2(n-1)}}{(n-1)!} = e^{\varsigma^2}.$$

Multiply this by the $e^{-\varsigma^2/2}$ factor to recover the full wavefunction, we find ψ diverges for large ξ as $+\varsigma^2/2$.

Actually we should have expected this for a general value of the energy, the Schrodinger equation has the solution $\approx Ae^{+\varsigma^2/2} + Be^{-\varsigma^2/2}$ at large distances, and only at certain energies does the coefficient A vanish to give a normalizable bound state wavefunction.

So how do we find the *nondiverging* solutions? It is clear that the infinite power series must be stopped! The key is in the recurrence relation.

If the energy satisfies

$$2\varepsilon = 2n+1, \quad n \text{ an integer,}$$

then h_{n+2} *and all higher coefficients vanish.* This requirement in fact *completely determines the polynomial* (except for an overall constant) because with $2\varepsilon = 2n+1$ the coefficients h_m for $m < n$ are determined by

$$h_{m+2} = \frac{(2m+1-2\varepsilon)}{(m+1)(m+2)}h_m = \frac{(2m+1-(2n+1))}{(m+1)(m+2)}h_m.$$

This n^{th} order polynomial is called a *Hermite polynomial* and written $H_n(\xi)$. The standard normalization of the Hermite polynomials $H_n(\xi)$ is to take the coefficient of the highest power ξ^n to be 2^n. The other coefficients then follow using the recurrence relation above, giving:

$$H_0(\xi) = 1,\ H_1(\xi) = 2\xi,\ H_2(\xi) = 4\xi^2 - 2, H_3(\xi) = 8\xi^3 - 12\xi, etc.$$

So the bottom line is that the wavefunction for the n^{th} excited state, having energy $\varepsilon = n + \frac{1}{2}$, is $\psi_n(\xi) = C_n H_n(\xi) e^{-\varsigma^2/2}$. It can be shown that. Using this, beginning with the ground state, one can easily convince oneself that the successive energy eigenstates each have one more node the n^{th} state has n nodes. This is also evident from numerical solution using the spreadsheet, watching how the wave function behaves at large x as the energy is cranked up.The spreadsheet can also be used to plot the wave function for large n, say $n = 200$. It is instructive to compare the probability distribution with that for a *classical* pendulum, one oscillating with fixed amplitude and observed many times at random intervals. For the pendulum, the probability peaks at the end of the swing, where the pendulum is slowest and therefore spends most time. The $n = 200$ distribution amplitude follows this pattern, but of course oscillates. However, in the large n limit these oscillations take place over undetectably small intervals.The *classical* pendulum when not at rest clearly has a time-dependent probability distribution it swings backwards and forwards. This means it *cannot* be in an eigenstate of the energy. In fact, the quantum state most like the classical is a *coherent state* built up of neighbouring energy eigenstates.

Operator Approach to the Simple Harmonic Oscillator

Having scaled the position coordinate x to the dimensionless $\xi \, by \, \xi = x / b = x\sqrt{m\omega / \hbar}$, let us also scale the momentum from p to $\pi = -id / d\xi$ (so $\pi = bp / \hbar = p / \sqrt{\hbar m\omega}$).

The Hamiltonian is

$$H = \frac{p^2 + m^2\omega^2 x^2}{2m} = \frac{\hbar\omega}{2}\left(\pi^2 + \xi^2\right).$$

Dirac had the brilliant idea of factorizing this expression: the obvious thought $\left(\xi^2 + \pi^2\right) = (\xi + i\pi)(\xi - i\pi)$ isn't quite right, because it fails to take account of the noncommutativity of the operators, but the symmetrical version

$$H = \frac{\hbar\omega}{4}[(\xi + i\pi)(\xi - i\pi) + (\xi - i\pi)(\xi + i\pi)]$$

is fine, and we shall soon see that it leads to a very easy way of finding the eigenvalues and operator matrix elements for the oscillator, far simpler than using the wave functions we found above. Interestingly, Dirac's factorization here of a second-order differential operator into a product of first-order operators is close to the idea that led to his most famous achievement, the Dirac equation, the basis of the relativistic theory of electrons, protons, etc.

To continue, we define new operators $a, a^\dagger$ by

$$a = \frac{\xi + i\pi}{\sqrt{2}} = \frac{1}{\sqrt{2\hbar m\omega}}(m\omega x + ip), a^\dagger = \frac{\xi - i\pi}{\sqrt{2}} = \frac{1}{\sqrt{2\hbar m\omega}}(m\omega x - ip).$$

From the commutation relation $[i\pi, \xi] = 1$ it follows that

$[a,a^{\dagger}]=1$ Therefore the Hamiltonian can be written:

$$H=\hbar\omega\left(a^{\dagger}a+\frac{1}{2}\right)=\hbar\omega\left(N+\frac{1}{2}\right),\text{ where } N=a^{\dagger}a.$$

Note that the operator N can only have *non-negative* eigenvalues, since

$$\langle\psi|N|\psi\rangle=\left\langle\psi\left|a^{\dagger}a\right|\psi\right\rangle=\langle\psi_a|\psi_a\rangle\geq 0.$$

Now

$$\left[N,a^{\dagger}\right]=a^{\dagger}aa^{\dagger}-a^{\dagger}a^{\dagger}a=a^{\dagger}\left[a,a^{\dagger}\right]=a^{\dagger}$$

Suppose N has an eigenfunction $|\nu\rangle$ with eigenvalue ν,

$$\mathrm{N}|\nu\rangle=\nu\,|\nu\rangle,$$

From the two equations above

$$Na^{\dagger}|\nu\rangle=a^{\dagger}N|\nu\rangle+a^{\dagger}|\nu\rangle=(\nu+1)a^{\dagger}|\nu\rangle$$

so $a^{\dagger}|\nu\rangle$ is an eigenfunction of N with eigenvalue $\nu+1$. Operating with again and again, we climb an infinite ladder of eigenstates equally spaced in energy is often termed a *creation operator*, since the quantum of energy added each time it operates is equivalent to an added photon in black body radiation (electromagnetic oscillations in a cavity). It is easy to check that the state $a|\nu\rangle$ is an eigenstate with eigenvalue $\nu-1$, provided it is nonzero, so the operator a takes us *down* the ladder. However, this cannot go on indefinitely we have established that N cannot have negative eigenvalues. We must eventually reach a state $|\nu\rangle$ for which $a|\nu\rangle=0$ the operator a *annihilates* the state. (At each step down, a annihilates one quantum of energy so a is often called an *annihilation* or *destruction* operator.)

Since the norm squared of $a|\nu\rangle$,

$$|a|\nu\rangle|^2=\left\langle\nu\left|a^{\dagger}a\right|\nu\right\rangle=\langle\nu|N|\nu\rangle=\nu|\langle\nu|\nu\rangle,$$

and since $\langle\nu|\nu\rangle>0$ for any nonvanishing state, it must be that *the lowest eigenstate* (the $|\nu\rangle$ for which $a|\nu\rangle$=0) *has* $\nu=0$. It follows that the ν 's on the ladder are *the positive integers*, so from this point on we relabel the eigenstates with n in place of ν.

That is to say, we have proved that the only possible eigenvalues of N are zero and the positive integers: 0, 1, 2, 3… . N is called the *number operator*: it measures the number of quanta of energy in the oscillator above the irreducible ground state energy (that is, above the 'zero-point energy' arising from the wave-like nature of the particle).

Since from above the Hamiltonian

$$H = h\omega\left(a^{\uparrow}a + \frac{1}{2}\right) = \hbar\omega\left(N + \frac{1}{2}\right)$$

the energy eigenvalues are

$$H|n\rangle = \left(n + \frac{1}{2}\right)\hbar\omega|n\rangle.$$

It is important to appreciate that Dirac's factorization trick and very little effort has given us *all* the eigenvalues of the Hamiltonian

$$H = \frac{h\omega}{2}\left(\pi^2 + \xi^2\right).$$

Contrast the work needed in this section with that in the standard Schrodinger approach. We have also established that the lowest energy state $|0\rangle$, having energy $\frac{1}{2}\hbar\omega$, must satisfy the first-order differential equation that is,

$$(\xi + i\pi)\Big|0 >= \left(\xi + \frac{d}{d\xi}\right)\psi_0(\xi) = 0.$$

The solution, unnormalized, is

$$\psi_0(\xi) = Ce^{-\varsigma^2/2}.$$

(In fact, we've seen this equation and its solution before: this was the condition for the 'least uncertain' wave function in the discussion of the Generalized Uncertainty Principle.)

We denote the *normalized* set of eigenstates $|0\rangle, |1\rangle, |2\rangle, \ldots. |n\rangle \ldots.$ with $\langle n|n\rangle = 1$. Now $a^{\uparrow}|n\rangle = C_n|n+1\rangle$ and C_n is easily found:

$$|C_n| = |C_n|^2\langle n+1|n+1\rangle = \left\langle n\middle|aa^{\uparrow}\middle|n\right\rangle = (n+1),$$

and $a^{\uparrow}|n\rangle = \sqrt{n+1}|n+1\rangle.$

Therefore, if we take the set of orthonormal states $|0\rangle, |1\rangle, |2\rangle, \ldots. |n\rangle \ldots..$ as the basis in the Hilbert space, the *only* nonzero matrix elements of $\langle n+1|a^{\uparrow}|n\rangle = \sqrt{n+1}$ are That is to say,

$$a^{\uparrow} = \begin{bmatrix} 0 & 0 & 0 & 0 & \ldots \\ \sqrt{1}0 & 0 & 0 & 0 & \ldots \\ 0 & \sqrt{2} & 0 & 0 & \ldots \\ 0 & 0 & \sqrt{3} & 0 & \ldots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

(The column vectors in the space this matrix operates on have an infinite number of elements: the lowest energy, the ground state component, is the entry at the *top* of the infinite vector so up the energy ladder is down the vector!)

The adjoint

$$a = \begin{bmatrix} 0 & \sqrt{1} & 0 & 0 & \ldots \\ 0 & 0 & \sqrt{2} & 0 & \ldots \\ 0 & 0 & 0 & \sqrt{3} & \ldots \\ 0 & 0 & 0 & 0 & \ldots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

So, $a|n\rangle = \sqrt{n}|n-1\rangle.$

For practical computations, we need to find the matrix elements of the position and momentum variables between the normalized eigenstates. Now

$$x = \sqrt{\hbar/2m\omega}\left(a^{\uparrow} + a\right),\ p = i\sqrt{m\omega\hbar/2}\left(a^{\uparrow} + a\right)$$

so

$$x = \sqrt{\hbar/2m\omega}\begin{bmatrix} 0 & \sqrt{1} & 0 & 0 & \ldots \\ \sqrt{1} & 0 & \sqrt{2} & 0 & \ldots \\ 0 & \sqrt{2} & 0 & \sqrt{3} & \ldots \\ 0 & 0 & \sqrt{3} & 0 & \ldots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

$$p = i\sqrt{m\omega\hbar/2}\begin{bmatrix} 0 & -\sqrt{1} & 0 & 0 & \cdots \\ \sqrt{1} & 0 & -\sqrt{2} & 0 & \cdots \\ 0 & \sqrt{2} & 0 & -\sqrt{3} & \cdots \\ 0 & 0 & \sqrt{3} & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

These matrices are, of course, Hermitian (not forgetting the *i* factor in *p*).

To find the matrix elements between eigenstates of any product of x's and p's, express all the x's and p's in terms of a's and $a^{\dagger}$'s, to give a sum of products of a's and $a^{\dagger}$'s. Each product in this sum can be evaluated sequentially from the right, because each a or $a^{\dagger}$ has only one nonzero matrix element when the product operates on one eigenstate.

Normalizing the Eigenstates in x*-space*

The normalized ground state wave function is

$$\psi_0(\xi) = C^{-\zeta^2/2} = \left(\frac{m\omega}{\pi\hbar}\right)e^{-m\omega x^2/2k},$$

where we have gone back to the x variable, and normalized using

$$\int_{-\infty}^{\infty} e^{-\alpha x^2}dx = \sqrt{\pi/a}.$$

To find the normalized wave functions for the higher states, they are first constructed formally by applying the creation operator $a^{\dagger}$ repeatedly on the ground state using $\langle n|a^{\dagger}|n-1\rangle\sqrt{n}$,

$$|n\rangle = \frac{a^{\dagger}}{\sqrt{n}}|n-1\rangle = \ldots\frac{\left(a^{\dagger}\right)^n}{\sqrt{n!}}|0\rangle.$$

Now

$$a^{\dagger} = \left(1/\sqrt{2}\right)(\xi - i\pi) = \left(1/\sqrt{2}\right)(\xi - d/d\xi),$$

so $$\psi_n(\xi) = \frac{\left(a^{\dagger}\right)^n}{\sqrt{n}}|0\rangle = \frac{1}{\sqrt{n!}}\left(\frac{1}{\sqrt{2}}\left(\xi - \frac{d}{d\xi}\right)\right)^n\left(\frac{m\omega}{\pi\hbar}\right)^{1/4}e^{-\zeta^2/2}.$$

We need to check that this expression is indeed the same as the Hermite polynomial wave function derived earlier, and to do that we need some further properties of the Hermite polynomials.

Some Properties of Hermite Polynomials

The mathematicians *define* the Hermite polynomials by:

$$H_n(\xi) = (-)^n e^{\varsigma^2} \frac{d^n}{d\xi^n} e^{-\varsigma^2}$$

so, $H_0(\xi) = 1$, $H_1(\xi) = 2\xi$, $H_2(\xi)$

$$= 4\xi^2 - 2,\ H_3(\xi) = 8\xi^3 - 12\xi,\ etc.$$

It follows immediately from the definition that the coefficient of the leading power is 2^n.

It is a straightforward exercise to check that H_n is a solution of the differential equation

$$\left(\frac{d^2}{d\xi^2} - 2\xi\frac{d}{d\xi} + 2n\right)H_n(\xi) = 0,$$

so these are indeed the same polynomials we found by the series solution of Schrodinger's equation earlier (recall the equation for the polynomial component of the wave function was

$$\frac{d^2h}{d\xi^2} - 2\xi\frac{dh}{d\xi} + (2n-1)h = 0,$$

with $2\omega = 2n+1$).

We have found $\psi_n(\xi)$ in the form

$$\psi_n(\xi) = \frac{1}{\sqrt{n!}}\left(\frac{1}{\sqrt{2}}\left(\xi - \frac{d}{d\xi}\right)\right)^n \left(\frac{m\omega}{\pi\hbar}\right)^{1/4} e^{-\varsigma^2/2}.$$

We shall now prove that the polynomial component is exactly equivalent to the Hermite polynomial as defined at the beginning of this section.

We begin with the operator identity:

$$\left(\xi - \frac{d}{d\xi}\right) = -e^{\varsigma^2/2}\frac{d}{d\xi}e^{-\varsigma^2/2}$$

Both sides of this expression are to be regarded as *operators*, that is, it is assumed that both are operating on some function $f(\xi)$.

Now take the n^{th} power of both sides: on the right, we find, for example,

$$\left(-e^{\zeta^2/2}\frac{d}{d\xi}e^{-\zeta^2/2}\right)^3 = (-)^3 e^{\zeta^2/2}\frac{d}{d\xi}e^{-\zeta^2}e^{\zeta^2/2}$$

$$(-)^3 e^{\zeta^2/2}\frac{d^3}{d\xi^3}e^{-\zeta^2/2}$$

since the intermediate exponential terms cancel against each other.

So:

$$\left(\xi-\frac{d}{d\xi}\right)^n = (-)^n e^{\zeta^2/2}\frac{d^n}{d\xi^n}e^{-\zeta^2/2}$$

and substituting this into the expression for $\psi_n(\xi)$ above,

$$\psi_n(\xi)=\frac{1}{\sqrt{2^n n!}}(-)^n\left(e^{\zeta^2/2}\frac{d^n}{d\xi^n}e^{-\zeta^2/2}\right)\left(\frac{m\omega}{\pi\hbar}\right)^{1/4}e^{-\zeta^2/2}$$

$$\frac{1}{\sqrt{2^n n!}}(-)^n\left(\frac{m\omega}{\pi\hbar}\right)^{1/4}e^{-\zeta^2/2}\left(e^{\zeta^2}\frac{d^n}{d\xi^n}e^{-\zeta^2}\right)$$

$$\frac{1}{\sqrt{2^n n!}}\left(\frac{m\omega}{\pi\hbar}\right)^{1/4}H_n(\zeta)e^{-\zeta^2/2},\ \text{with } \xi=\sqrt{\frac{m\omega}{\hbar}}x.$$

This established the equivalence of the two approaches to Schrodinger's equation for the simple harmonic oscillator, and provides us with the overall normalization constants without doing integrals. (The expression for $\psi_n(\xi)$ above satisfies $\int|\psi_n|^2\,dx=1$.)

Use $H_n(\xi)=(-)^n e^{\zeta^2}\frac{d^n}{d\xi^n}e^{-\zeta^2}$ to prove:

(a) The coefficient of ξ^n is 2^n.

(b) $H_n'(\xi)=2nH_{n-1}(\xi)$

(c) $H_{n+1}(\xi) = 2\xi H_n(\xi) - 2nH_{n-1}(\xi)$

(d) $\int_{-\infty}^{\infty} e^{-\xi^2} H_n^2(\xi)d\xi = 2^n n!\sqrt{\pi}$

(Hint: rewrite as $\int_{-\infty}^{\infty} H_n(\xi)(-)^n \frac{d^n}{d\xi^n} e^{-\xi^2} d\xi$, then integrate by parts n times, and use (a).)

(e) $\int_{-\infty}^{\infty} e^{-\xi^2} H_n(\xi)H_m(\xi)d\xi = 0$ for $m \neq n$.

It's worth doing these exercises to become more familiar with the Hermite polynomials, but in evaluating matrix elements (and indeed in establishing some of these results) it is almost always far simpler to work with the creation and annihilation operators.

Use the creation and annihilation operators to find $\langle n|x^4|n\rangle$. This matrix element is useful in estimating the energy change arising on adding a small nonharmonic potential energy term to a harmonic oscillator.

7

Time-dependent Wave Functions

The set of normalized eigenstates $|0\rangle, |1\rangle, |2\rangle, ... |n\rangle ...$ discussed above are of course solutions to the time-independent Schrodinger equation, or in ket notation eigenstates of the Hamiltonian $H|n\rangle = \left(n + \frac{1}{2}\right)\hbar\omega|n\rangle$. Putting in the time-dependence explicitly, $|n,t\rangle = e^{iH/k}|n,t=0\rangle = e^{-i\left(n+\frac{1}{2}\right)\omega t}|n\rangle$. It is necessary to include the time dependence when dealing with a state which is a superposition of states of different energies, such as

$$\left(1/\sqrt{2}\right)\left(|0\rangle + |1\rangle\right),$$

which then becomes.

$$\left(1/\sqrt{2}\right)\left(e^{-i\omega t/2}|0\rangle\right) + \left(e^{-3i\omega t/2}|1\rangle\right).$$

Expectation values of combinations of position and/or momentum operators in such states are best evaluated by expressing everything in terms of annihilation and creation operators.

Solving Schrodinger's Equation in Momentum Space

In the lecture on Function Spaces, we established that the basis of $|x\rangle$ states (eigenstates of the position operator) and that of $|k\rangle$ states (eigenstates of the momentum operator) were both complete bases in Hilbert space (physicist's definition) so we could work equally well with either from a formal point of view. Why then do we almost always work in x-space? Well, probably because we live in x-space, but there's another reason. The momentum operator in the x-space representation is $p = -i\hbar d/dx$, so Schrodinger's equation, written

$(p^2/2m + V(x))\psi(x) = E\psi(x)$, with p in operator form, is a second-order differential equation. Now consider what happens to Schrodinger's equation if we work in p-space. Since the operator identity $[x,p] = i\hbar$ is true regardless of representation, we must have $x = i\hbar d/dp$. So for a particle in a potential $V(x)$, writing Schrodinger's equation in p-space we are confronted with the nasty looking operator $V(i\hbar d/dp)$! This will produce a differential equation in general a lot harder to solve than the standard x-space equation so we stay in x-space.

But there are two potentials that can be handled in momentum space: first, for a *linear* potential $V(x) = -Fx$, the momentum space analysis is actually easier it's just a first-order equation. Second, for a particle in a *quadratic* potential a simple harmonic oscillator the two approaches yield the *same* differential equation. That means that the eigenfunctions in momentum space (scaled appropriately) must be *identical* to those in position space. The simple harmonic eigenfunctions are their own Fourier transforms!

Time-dependent Solutions: Propagators and Representations

We've spent most of the course so far concentrating on the eigenstates of the Hamiltonian, states whose time-dependence is merely a changing phase. We did mention much earlier a superposition of two different energy states in an infinite well, resulting in a wave function sloshing backwards and forwards. It's now time to cast the analysis of time dependent states into the language of bras, kets and operators. We'll take a time-independent Hamiltonian H, with a complete set of orthonormalized eigenstates, and as usual

$$i\hbar\frac{\partial\psi(x,t)}{\partial t} = -\frac{\hbar^2}{2m}\frac{\partial^2\psi(x,t)}{\partial x^2} + V(x)\psi(x,t).$$

Or, as we would now write it

$$i\hbar\frac{\partial}{\partial t}|\psi(x,t)\rangle = H|\psi(x,t)\rangle.$$

Since H is itself time independent, this is very easy to integrate!

$$\left|\psi(x,t)\right\rangle = e^{-iH(t-t_0)/k}\left|\psi(x,t_0)\right\rangle.$$

The exponential operator that generates the time-dependence is called the *propagator*, because it describes how the wave propagates from its initial configuration, and is usually denoted by *U*:

$$\left|\psi(x,t)\right\rangle = U(t-t_0)\left|\psi(x,t_0)\right\rangle.$$

It's appropriate to call the propagator *U*, because it's a unitary operator:

$$U(t-t_0) = e^{-iH(t-t_0)/k},$$

so $U^{\uparrow}(t-t_0) = e^{-iH'(t-t_0)/k} = e^{-iH(t-t_0)/k} = U^{-1}(t-t_0).$

Since *H* is Hermitian, *U* is unitary. It immediately follows that

$$\left\langle\psi(x,t)\middle|\psi(x)\right\rangle = \left\langle\psi(x,t_0)\middle|U^{\uparrow}U(t-t_0)\middle|\psi(x,t_0)\right\rangle$$

$$\left\langle\psi(x,t_0)\middle|\psi(x,t_0)\right\rangle$$

the norm of the ket vector is conserved, or, translating to wave function language, a wave function correctly normalized to give a total probability of one stays that way. (This can also be proved from the Schrodinger equation, of course, but this is quicker.)

This is all very succinct, but unfortunately the exponential of a second-order differential operator doesn't sound too easy to work with. Recall, though, that any function of a Hermitian operator has the same set of eigenstates as the original operator. This means that the eigenstates of $e^{-iH(t-t_0)/k}$ are the *same* as the eigenstates of *H*, and if $H\left|\psi_n\right\rangle = E_n\left|\psi_n\right\rangle$, then

$$e^{-iH(t-t_0)/k}\left|\psi_n\right\rangle = e^{-iE_n(t-t_0)/k}\left|\psi_n\right\rangle$$

This is of course nothing but the time dependent phase factor for the eigenstates we found before and, as before, to find the time dependence of any general state we must express it as a superposition of these eigenkets, each having its own time dependence. But how do we do that in the operator language? Easy: we simply insert an identity operator, the one constructed from the complete set of eigenkets, thus:

$$\left|\psi(t)\right\rangle = e^{-H(t-t_0)/k}\sum_{n-1}^{\infty}\left|\psi_n\right\rangle\left\langle\psi_n\middle|\psi(t_0)\right\rangle$$

$$\sum_{n-1}^{\infty} e^{-\imath E_n(t-t_0)/k}\left|\psi_n\right\rangle\left\langle\psi_n\middle|\psi(t_0)\right\rangle.$$

Staring at this, we see that it's just what we had before: at the initial time' $t = t_0$, the wave function can be written as a sum over the eigenkets:

$$\left|\psi(t_0)\right\rangle = \sum\left|\psi_n(t_0)\right\rangle\left\langle\psi_n(t_0)\middle|\psi(t_0)\right\rangle = \sum c_n\left|\psi_n(t_0)\right\rangle$$

with $c_n = \left\langle\psi_n\middle|\psi\right\rangle, \sum\left|c_n\right|^2 = 1$, and the usual generalization for continuum eigenvalues, and the time development is just given by inserting the phases:

$$\left|\psi(t)\right\rangle = \sum c_n e^{iE_n(t-t_0)/k}\left|\psi_n(t_0)\right\rangle.$$

The expectation value of the energy E in $\left|\psi\right\rangle$,

$$\left\langle E\right\rangle = \left\langle\psi\middle|H\middle|\psi\right\rangle = \sum\left|c_n\right|^2 E_n$$

and is (of course) time independent.

The expectation value of the particle position x is

$$\left\langle\psi(t)\middle|x\middle|\psi(t)\right\rangle = \sum_{n,m} c_n^* c_m e^{i(E_n-E_m)(t-t_0)/k}\left\langle\psi_n(t_0)\middle|x\middle|\psi_m(t_0)\right\rangle$$

and is *not* in general time-independent. (It *is* real, of course, on adding the (n,m) term to the (m,n) term.) This analysis is only valid for a *time-independent* Hamiltonian. The important extension to a system in a time-dependent external field, such as an atom in a light beam.

The Free Particle Propagator

To gain some insight into what the propagator U looks like, we'll first analyse the case of a particle in one dimension with no potential at all.

We'll also take $t_0 = 0$ to make the equations less cumbersome.

For a free particle in one dimension $E = p^2/2m = \hbar^2k^2/2m$ the energy eigenstates are also momentum eigenstates, we label them $\left|k\right\rangle$, so

$$U(t)=e^{-iHt/k}=\int_{-\infty}^{\infty}e^{-iHt/k}\frac{dk}{2\pi}|k\rangle|k\rangle=\int_{-\infty}^{\infty}e^{-ikk^2t/2m}\frac{dk}{2\pi}|k\rangle\langle k|$$

Let's consider what seems the simplest example.

Suppose that at $t = t_0 = 0$, a particle is at x_0: $\psi(x,t=0)=\delta(x-x_0)=|x_0\rangle$: what is the probability amplitude for finding it at x at a later time t?

$$\langle x|U(t,0)|x_0\rangle=\int_{-\infty}^{\infty}e^{-kk^2t/2m}\frac{dk}{2\pi}\langle x|k\rangle\langle k|x_0\rangle$$

$$=\int_{-\infty}^{\infty}e^{-kk^2t/2m}\frac{dk}{2\pi}^{-ik(x_0-x)}$$

$$=\sqrt{\frac{m}{2\pi\hbar it}}e^{im(x_0-x)^2/\kappa}$$

using the standard identity for Gaussian integrals,

$$\int_{-\infty}^{\infty}dke^{-ak^2+bk}=\sqrt{\frac{\pi}{a}}e^{b^2/4a}.$$

On examining the above expression, though, it turns out to be nonsense! Noting that the term in the exponent is pure imaginary, $|\psi(x,t)|^2=m/2\pi\hbar t$ independent of x! This particle apparently instantaneously fills all of space, but then its probability dies away as $1/t$.

Notice first that $|\psi(x,t)|^2$ is constant throughout space. This means that the normalization, $\int|\psi(x,t)|^2dx=\infty$! And, as we've seen above, the normalization stays constant in time the propagator is unitary. Therefore, our initial wave function must have had infinite norm. That's exactly right we took the initial wave function $\psi(x,t=0)=\delta(x-x_0)=|x_0\rangle$.

Think of the δ-function as a limit of a function equal to $1/\Delta$ over an interval of length Δ, with Δ going to zero, and it's clear the normalization goes to infinity as $1/\Delta$. This is not a meaningful wave function for a particle. Recall that continuum kets like $|x_0\rangle$ are normalized by $\langle x|x'\rangle = \delta(x-x')$, they do *not* represent wave functions individually normalizable in the usual sense. The only meaningful wave functions are *integrals over a range* of such kets, such as $\int dx\psi(x)|x\rangle$. In an integral like this, notice that states $|x\rangle$ within some tiny x-interval of length δx, say, have total weight $\psi(x)\delta x$, which goes to zero as δx is made smaller, but by writing $\psi(x,t=0)\delta(x-x_0) = |x_0\rangle$ we took a single such state and gave it a finite weight. This we can't do.

Of course, we do want to know how a wave function initially localized *near* a point develops. To find out, we must apply the propagator to a legitimate wave function one that is normalizable to begin with. The simplest "localized particle" wave function from a practical point of view is a Gaussian wave packet,

$$\psi(x',0) = e^{ip_0x'/k}\frac{e^{-x'^2/2d^2}}{\left(\pi d^2\right)^{1/4}}.$$

We used Δ in place of here.

The wave function at a later time is then given by the operation of the propagator on this initial wave function:

$$\psi(x,t) = \int U(x,t;x',0)e^{ip_0x'/k}\frac{e^{-x'^2/2d^2}}{\left(\pi d^2\right)^{1/4}}dx'$$

$$= \sqrt{\frac{m}{2\pi\hbar it}}\int e^{im(x-x')^2/2kt}e^{ip_0x'/k}\frac{e^{-x'^2/2d^2}}{\left(\pi d^2\right)1/4}dx'.$$

Note first that since this is just $|\psi(x,t)\rangle = U(t)|\psi(x,t=0)\rangle$ written explicitly in terms of Schrodinger wave functions,

$$\psi(x,t) = \int U(x,t;x',0)\psi(x',0)dx'$$

it is evident that $U(x.t;\ x'\ 0) \to \delta(x-x')$ as $t \to 0$. This is just equivalent to the operator statement that $e^{-iHt/k} \to 1$, the unit operator, as $t \to 0$.

The integral over x^2 is just another Gaussian integral, so we use the same result,

$$\int_{-\infty}^{\infty} dx' e^{-ax'^2+bx'} = \sqrt{\frac{\pi}{a}} e^{b^2/4a} .$$

Looking at the expression above, we can see that

$$b = -\frac{im}{\hbar t}.\left(x - \frac{p_0 t}{m}\right),\ a = \frac{1}{2d^2} - \frac{im}{2\hbar t}.$$

This gives

$$\psi(x,t) = \frac{\pi^{-1/4}}{\sqrt{d\left(1 + \frac{i\hbar t}{md^2}\right)}} \exp\left(\frac{imx^2}{2\hbar t}\right) \exp\left(-\frac{\frac{im}{\hbar t}\left(x - \frac{p_0 t}{m}\right)^2}{2\left(1 + \frac{i\hbar t}{md^2}\right)}\right)$$

where the second exponential is the term $e^{b^2/4a}$. As written, the small t limit is not very apparent, but some algebraic rearrangement yields:

$$\psi(x,t) = \frac{\pi^{-1/4}}{\sqrt{d\left(1 + i\hbar t/md^2\right)}} \exp\left(-\frac{\left(x - p_0 t/m^2\right)}{2d^2\left(1 + i\hbar t/md^2\right)}\right)$$

$$\exp\left(\frac{ip_0}{\hbar}\left(x - p_0 t/2m\right)\right)$$

Written this way, it is evident that the expression *goes to the initial wave packet* for t going to zero, as of course it must.

Although the phase in the above expression for $\psi(x,t)$ has contributions from all three terms, the main phase oscillation is in the third term, and one can see the phase velocity is one-half the group velocity, as discussed earlier.

The resulting *probability density*:

$$\left|\psi(x,t)\right|^2 = \frac{1}{\sqrt{\pi\left(d^2 + \hbar^2 t^2 / m^2 d^2\right)}}.\exp^{-\frac{\left(x - p_0 t/m\right)^2}{d^2 + k^2 t^2/m^2 a^2}}$$

This is a Gaussian wave packet, having a width which goes as $\hbar t / md$ for large times, where d is the width of the initial packet in x-space so $\hbar / md$ is the spread in *velocities* $\langle \Delta v \rangle$ within the packet, hence the gradual spreading in $\langle \Delta v \rangle t$ x-space.

It's amusing to look at the limit of this as the width d of the initial Gaussian packet goes to zero, and see how that relates to our δ-function result. Suppose we are at distance x from the origin, and there is initially a Gaussian wave packet centered at the origin, width $d << x$. At time $t \sim mxd / \hbar$, the wave packet has spread to x and has $|\psi(x,t)|^2$ of order $1/x$ at x. Thereafter, it continues to spread at a linear rate in time, so locally $|\psi(x,t)|^2$ must decrease as $1/t$ to conserve probability. In the δ-function limit $d \to 0$, the wave function instantly spreads through a huge volume, but then goes as $1/t$ as it spreads into an even huger volume.

Schrodinger and Heisenberg Representations

Assuming a Hamiltonian with no explicit time dependence, the time-dependent Schrodinger equation has the form

$$i\hbar \frac{\partial}{\partial t} |\psi(x,t)\rangle = H |\psi(x,t)\rangle$$

and as discussed above, the formal solution can be expressed as:

$$|\psi(x,t)\rangle = e^{-iht/k} |\psi(x,t=0)\rangle.$$

Now, any measurement on a system amounts to measuring a matrix element of an operator between two states (or, more generally, a function of such matrix elements).

In other words, the *physically significant time dependent quantities* are of the form

$$\langle \varphi(t) | A | \psi(t) \rangle = \langle \varphi(0) | e^{iHt/k} A e^{iHt/k} | \psi(0) \rangle$$

where A is an operator, which we are assuming has no explicit time dependence.

So in this Schrodinger representation, the time dependence of the measured value of an operator like x or p comes about because we measure the matrix element of an unchanging operator between bras and kets that are changing in time.

Heisenberg took a different approach: he assumed that the ket describing a quantum system did *not* change in time, it remained at $|\psi(0)\rangle$, but the *operators* evolved according to:

$$A_H(t) = e^{iHt/k} A_H(0) e^{-iHt/k}.$$

Clearly, this leads to the same physics as before. The equation of motion of the operator is:

$$i\hbar \frac{dA_H(t)}{dt} = [A_H(t), H].$$

The Hamiltonian itself does not change in time energy is conserved, or, to put it another way, H commutes with $e^{-iHt/k}$. But for a nontrivial Hamiltonian, say for a particle in one dimension in a potential,

$$H = p^2/2m + V(x)$$

the separate components *will* have time-dependence, parallel to the classical case: the *kinetic* energy of a swinging pendulum varies with time. (For a particle in a potential in an energy eigenstate the expectation value of the kinetic energy is constant, but this is not the case for any other state, that is, for a superposition of different eigenstates.) Nevertheless, the *commutator* of x, p will be time-independent:

$$[X_h(t), p_H(t)] = e^{iHt/k}[x_H(0), p_H(0)]e^{iHt/k}$$

$$= e^{iHt/k} i\hbar e^{-iHt/k} = i\hbar$$

(The Heisenberg operators are identical to the Schrodinger operators at $t = 0$.)

Applying the general commutator result

$$[A, BC] = [A, B]C + B[A, C],$$

$$\left[x_H(t), \frac{p^2{}_H(t)}{2m}\right] = \frac{i\hbar p_H(t)}{m}$$

so,

$$\frac{dx_H(t)}{dt} = \frac{p_H(t)}{m}$$

and since $[x_H(t), p_H(t)] = i\hbar$, $p_H(t) = -i\hbar d/dx_H(t)$,

$$\frac{dp_H(dt)}{dt} = \frac{1}{i\hbar}\left[p_H(t), V(x_H(t))\right] - \nabla V(x_H(t)).$$

This result could also be derived by writing $V(x)$ as an expansion in powers of x, then taking the commutator with p.

Notice from the above equations that the operators in the Heisenberg Representation obey the *classical* laws of motion! *Ehrenfest's Theorem*, that the *expectation values* of operators in a quantum state follow the classical laws of motion, follows immediately, by taking the expectation value of both sides of the operator equation of motion in a quantum state.

Simple Harmonic Oscillator in the Heisenberg Representation

For the simple harmonic oscillator, the equations are easily integrated to give:

$$x_H(t) = x_H(0)\cos\omega t + \left(p_H(0)/m\omega\right)\sin\omega t$$

$$p_H(t) = p_H(0)\cos\omega t - m\omega x_H(0)\sin\omega t.$$

We have put in the $_H$ subscript to emphasize that these are operators. It is usually clear from the context that the Heisenberg representation is being used, and the subscript $_H$ may be safely omitted.

The time-dependence of the *annihilation operator a* is:

$$a(t) = e^{\imath Ht/k} a(0) e^{-\imath Ht/k}$$

with

$$H = \hbar\omega\left(a^{\uparrow}(t)a(t) + \tfrac{1}{2}\right).$$

Note again that although H is itself time-independent, it is necessary to include the time-dependence of individual operators within H.

$$i\hbar\frac{d}{dt}a(t) = \left[a(t), H\right] = \hbar\omega\left[a(t), a^{\uparrow}(t)a(t)\right]$$
$$= \hbar\omega\left[a(t), a^{\uparrow}(t)\right]a(t) = \hbar\omega a(t)$$

so,

$$a(t) = a(0)e^{-i\omega t}.$$

Actually, we could have seen this as follows: if $|n\rangle$ are the energy eigenstates of the simple harmonic oscillator,

$$e^{-iHt/k}|n\rangle = e^{-inkωt/k}|n\rangle = e^{-inωt}|n\rangle.$$

Now the *only* nonzero matrix elements of the annihilation operator $\hat{a}$ between energy eigenstates are of the form

$$\langle n-1|a(t)|n\rangle = |n-1\rangle e^{iHt/k}a(0)e^{-Ht/k}|n\rangle$$

$$= e^{iω(n-1)t}\langle n-1|a(0)|n\rangle e^{-ωnt} = \langle n-1|a(0)|n\rangle e^{-iωt}$$

Since this time-dependence is true of *all* energy matrix elements (trivially so for most of them, since they're identically zero), and the eigenstates of the Hamiltonian span the space, it is true as an *operator* equation.

Evidently, the expectation value of the operator $a(t)$ in any state goes clockwise in a circle centered at the origin in the complex plane. That this is indeed the classical motion of the simple harmonic oscillator is confirmed by recalling the definition $a = \frac{\xi + i\pi}{\sqrt{2}} = \frac{1}{\sqrt{2\hbar m\omega}}(m\omega x + ip)$, so the complex plane corresponds to the $(m\omega x, p)$ phase space discussed near the beginning of the lecture on the Simple Harmonic Oscillator. The time-dependence of the creation operator is just the adjoint equation: $a^{\dagger}(t) = a^{\dagger}(0)e^{i\omega t}$.

8

Simple Harmonic Oscillator

Consider a macroscopic simple harmonic oscillator, and to keep things simple assume there are no interactions with the rest of the universe. We know how to describe the motion using classical mechanics: for a given initial position and momentum, classical mechanics correctly predicts the future path, as confirmed by experiments with real (admittedly not perfect) systems. But from the Hamiltonian we could also write down Schrodinger's equation, and from that predict the future behaviour of the system. Since we already know the answer from classical mechanics and experiment, quantum mechanics must give us the same result in the limiting case of a large system.

It is a worthwhile exercise to see just how this happens. Evidently, we cannot simply follow the classical method of specifying the initial position and momentum the uncertainty principle won't allow it. What we can do, though, is to take an initial state in which the position and momentum are specified *as precisely as possible*. Such a state is called a *minimum uncertainty* state.

In fact, the *ground state* of a simple harmonic oscillator *is* a minimum uncertainty state. This is not too surprising it's just a localized wave packet centered at the origin. The system is as close to rest as possible, having only zero-point motion. What *is* surprising is that there are excited states of the pendulum in which this ground state wave packet swings backwards and forwards indefinitely, a quantum realization of the classical system, and the wave packet is always one of minimum uncertainty. Recall that this *doesn't* happen for a *free* particle on a line in that case, an initial minimal uncertainty wave packet spreads out because the different momentum components move at different speeds. But for the oscillator, the potential somehow keeps the wave packet together, a minimum uncertainty wave packet at all times. These remarkable quasi-classical states are called *coherent states*, and were discovered by Schrodinger himself. They are important in many quasi-classical contexts, including laser radiation.

Our task here is to construct and analyse these coherent states and to find how they relate to the usual energy eigenstates of the oscillator.

Classical Mechanics of the Simple Harmonic Oscillator

To define the notation, let us briefly recap the dynamics of the *classical* oscillator: the constant energy is

$$E=\frac{p^2}{2m}+\frac{1}{2}kx^2$$

or

$$p^2+(m\omega x)^2=2mE,\ \omega=\sqrt{k/m}.$$

The classical motion is most simply described in *phase space*, a two-dimensional plot in the variables $(m\omega x, p)$. In this space, the point $(m\omega x, p)$ corresponding to the position and momentum of the oscillator at an instant of time moves as time progresses at constant angular speed in a clockwise direction around the circle of radius $\sqrt{2mE}$ centred at the origin.

Phase space is usually defined in terms of the variables (x,p), but in describing the simple harmonic oscillator, the pair $(m\omega x, p)$ are more convenient.

This motion is elegantly described by regarding the two-dimensional phase space as a complex plane, and defining the dimensionless complex variable

$$z=\frac{m\omega x+ip}{\sqrt{2\hbar m\omega}}.$$

The time evolution in phase space is simply $z(t)=z_0e^{-i\omega t}$.

The particular choice of (quantum!) scaling factor in defining z amounts to defining the unit of energy as $\hbar\omega$, the natural unit for the oscillator: it is easy to check that if the classical energy $E = \left(n + \frac{1}{2}\right)\hbar\omega$ then the dimensionless $|z|^2$ is simply the number (which is of course very large, so the ½ is insignificant).

Minimum Uncertainty Wave Packets

The Generalized Uncertainty Principle that any minimum uncertainty one-dimensional wave function (so $\Delta p \times \Delta x = \hbar/2$) for a particle must satisfy the linear differential equation (here $\hat{p} = -i\hbar d/dx$)

$$\left(\hat{p} - \langle p\rangle\right)\psi(x) = \lambda\left(\hat{x} - \langle x\rangle\right)\psi(x)$$

where $\langle x\rangle, \langle p\rangle, \lambda$ are constants, and λ is *pure imaginary*. The equation is easy to solve: any minimum uncertainly one-dimensional wave function is a Gaussian wave packet, having expectation value of momentum $\langle p\rangle$, centered at $\langle x\rangle$ and having width $\left(\Delta x^2\right) = -\hbar/2i\lambda$. ($\Delta x$ is defined for a state $|\psi\rangle$ by $\left(\Delta x^2\right) = |\psi\rangle\left(x - \langle x\rangle\right)^2|\psi\rangle$.) That is to say, the minimum uncertainly solution is:

$$\psi(x) = Ce^{\imath\langle p\rangle x/k}e^{\imath\lambda\left(x-\langle x\rangle\right)^2/2k} = Ce^{\imath\langle p\rangle x/k}e^{-\left(x-\langle x\rangle\right)^2/4k(\Delta X)^2}$$

with C the normalization constant.

In fact, the simple harmonic oscillator *ground state* $\psi_0(x) = \left(\frac{m\omega}{\pi\hbar}\right)^{1/4} e^{-m\omega x^2/2k}$ is just such a minimum uncertainty state, with

$$\lambda - im\omega,\ \langle x\rangle = \langle p\rangle = 0;\ (\Delta x)^2 = \frac{\hbar}{2m\omega},\ (\Delta p)^2 = \frac{\hbar m\omega}{2}, \Delta p \times \Delta x = \frac{\hbar}{2}.$$

Furthermore, it is easy to see that the *displaced* ground state $\psi_0\left(x - \langle x_0\rangle\right) = Ce^{-m\omega\left(x-\langle x_0\rangle\right)^2/2k}$ (writing the normalization constant $\left(m\omega/\pi\hbar\right)^{1/4} = c$) must also be a minimum uncertainty state, *with the same* $\lambda = im\omega$. Of course, in contrast to the ground state, this displaced state is no longer an eigenstate of the Hamiltonian, and any such initial state will change with time.

(Both these states have the same spread in x-space $(\Delta x)^2 = \hbar / 2m\omega$, and the same spread in p-space, the only difference in the p direction being a phase factor $e^{ip\langle x_0\rangle / k}$ for the displaced state.)

What about the higher eigenstates of the oscillator Hamiltonian? They are not minimally uncertain states for the n^{th} state, $\Delta p \times \Delta x = n\hbar / 2$,

as is easily checked using

$$\frac{1}{2}(\Delta p)^2 / 2m = \frac{1}{2} k (\Delta x)^2 \sim \frac{1}{2} n\hbar\omega.$$

So, if we construct a minimally uncertain higher energy state, it will *not* be an eigenstate of the Hamiltonian.

Time Development of a Coherent State: The Role of the Annihilation Operator

In this section, we shall establish a remarkable connection between minimally uncertain oscillator states and the annihilation operator, then use properties of that operator to find the time-development of the minimally uncertain states.

Suppose that at $t = 0$ the oscillator wave function is the minimum uncertainty state

$$\psi(x, t=0) = Ce^{i\langle p_0\rangle x / k} e^{i\lambda (x - \langle x_0\rangle)^2 / 2k} = Ce^{i\langle p_0\rangle x / h} e^{-m\omega (x - \langle x_0\rangle)^2 / 2k}$$

centred at $\langle p_0\rangle, m\omega\langle x_0\rangle$ in phase space (as defined above for the classical oscillator), and with $\lambda = im\omega$ to give it the same spatial extent as the ground state.

From the preceding section, this $\psi(x,0)$ satisfies the minimum uncertainty equation

$$(\hat{p} - \langle p_0\rangle)\psi(x,0) = im\omega(\hat{x} - \langle x_0\rangle)\psi(x,0).$$

Rearranging this equation (and multiplying by $-i$) shows it in a different light:

$$(m\omega\hat{x} + i\hat{p})\psi(x,0) = (m\omega\langle x_0\rangle + i\langle p_0\rangle)\psi(x,0).$$

This is an eigenvalue equation! The wave packet $\psi(x,0)$ is an eigenstate of the operator $(m\omega\hat{x} + i\hat{p})$ with eigenvalue

$(m\omega\langle x_0\rangle + i\langle p_0\rangle)$. It is *not*, of course, an eigenstates of either $\hat{p}$ or $\hat{x}$ taken individually.

Furthermore, the operator $(m\omega\hat{x} + i\hat{p})$ is just a constant times the annihilation operator recall

$$\hat{a} = \frac{1}{\sqrt{2\hbar m\omega}}(m\omega\hat{x} + i\hat{p}).$$

Therefore, this minimally uncertain initial wave packet $\psi(x,0)$ is an eigenstate of the annihilation operator $\hat{a}$, with eigenvalue $(m\omega\langle x_0\rangle + i\langle p_0\rangle)/\sqrt{2\hbar m\omega}$. By the way, it's ok for to have complex eigenvalues, because $\hat{a}$ is not a hermitian operator.

We can now make the connection with the complex plane representation of the classical operator: the eigenvalue

$$(m\omega\langle x_0\rangle + i\langle p_0\rangle)/\sqrt{2\hbar m\omega}$$

is precisely the dimensionless complex parameter z_0!

This means that if we have a minimal uncertainty oscillator wave packet

$$\psi(x,t=0) = Ce^{i\langle p_0\rangle x/k}e^{-m\omega}\left(x\langle x_0\rangle\right)^2/2k$$

having the same spatial extent as the ground state, centred at $m\omega\langle x_0\rangle, \langle p_0\rangle$ in phase space, and we write

$$z_0 = \frac{m\omega\langle x_0\rangle + i\langle p_0\rangle}{\sqrt{2\hbar m\omega}}$$

then

$$\hat{a}\psi(x,t=0) = z_0\psi(x,t=0).$$

That is to say, the eigenstates of the annihilation operator $\hat{a}$ are all those minimal uncertainty wave packets that have the same spatial width as the oscillator ground state.

Turning now to the *time development* of the state, it is convenient to use the ket notation

$$|\psi(x,t=0)\rangle = |\langle x_0\rangle, \langle p_0\rangle\rangle$$

with $|x,p\rangle$ denoting a minimum uncertainly wave packet having those expectation values of position and momentum.

The time development of the ket, as usual, is given by

$$\left|\psi(x,t)\right\rangle = e^{-iHt/k}\left|\langle x_0\rangle,\langle p_0\rangle\right\rangle.$$

We shall show that $\left|\psi(x,t)\right\rangle$ *remains* an eigenstate of the annihilation operator *for all times t*: it therefore continues to be a minimum uncertainty wave packet!

The key point in establishing this is that the annihilation operator itself has a simple time development in the Heisenberg representation,

$$\hat{a}(t) = e^{iHt/k}\hat{a}e^{-iHt/k} = \hat{a}e^{-i\omega t}.$$

To prove this, consider the matrix elements of $\hat{a}(t)$ between any two eigenstates $|n\rangle$ of the Hamiltonian

$$H|n\rangle = \left(n+\frac{1}{2}\right)\hbar\omega|n\rangle$$

so,

$$\langle m|\hat{a}(t)|n\rangle = e^{i\left(m+\frac{1}{2}\right)k\omega t/k}\langle m|\hat{a}|n\rangle e^{i\left(n+\frac{1}{2}\right)k\omega t/k} = \langle n-1|\hat{a}|n\rangle e^{-i\omega t}.$$

Since the only nonzero matrix elements of the annihilation operator $\langle m|\hat{a}(t)|n\rangle$ are for m=n-1, and the energy eigenstates form a complete set, this simple time dependence is true as an *operator* equation

$$\hat{a}(t) = e^{iHt/k}\hat{a}e^{-iHt/k} = \hat{a}e^{-i\omega t}.$$

It is now easy to prove that

$$\left|\psi(x,t)\right\rangle = e^{-iHt/k}\left|\langle x_0\rangle,\langle p_0\rangle\right|$$

is always an eigenstate of $\hat{a}$:

$$\hat{a}\left|\psi(x,t)\right\rangle = \hat{a}e^{-iHt/k}\left|\langle x_0\rangle,\langle p_0\rangle\right\rangle$$

$$= e^{-iHt/k}\left(e^{iHt/k}\hat{a}e^{-iHt/k}\right)\left|\langle x_0\rangle,\langle p_0\rangle\right\rangle$$

$$= e^{iHt/k}e^{-i\omega t}\hat{a}\left|\langle x_0\rangle,\langle p_0\rangle\right\rangle$$

$$= e^{iHt/k}e^{-i\omega t}\left(m\omega\langle x_0\rangle + i\langle p_0\rangle\right)/\sqrt{2\hbar m\omega}\left|\langle x_0\rangle,\langle p_0\rangle\right\rangle$$

$$= (e^{-i\omega t}\left(m\omega\langle x_0\rangle + i\langle p_0\rangle\right)/\sqrt{2\hbar m\omega})\left|\psi(x,t)\right\rangle.$$

Therefore the annihilation operator, which at t = 0 had the eigenvalue

$$z_0 = \left(m\omega\langle x_0\rangle + i\langle p_0\rangle\right)/\sqrt{2\hbar m\omega},$$

corresponding to a minimal wave packet centred at $\left(m\omega\langle x_0\rangle, \langle p_0\rangle\right)$ in phase space, evolves in time t to another minimal packet (because it's still an eigenstate of the annihilation operator), and writing

$$\left|\langle x(t)\rangle, \langle p(t)\rangle\right\rangle = e^{-iHt/k}\left|\langle x_0\rangle, \langle p_0\rangle\right\rangle,$$

the new eigenvalue of $\hat{a}$

$$z(t) = \frac{\left(m\omega\left\langle x(t) + i\langle p(t)\rangle\right\rangle\right)}{\sqrt{2\hbar m\omega}} = \frac{\left(m\omega\langle x_0\rangle + i\langle p_0\rangle\right)}{\sqrt{2\hbar m\omega}}e^{-i\omega t} = z(0)e^{-i\omega t}$$

Therefore, the centre of the minimal wave packet in phase space follows the classical path in time. This is made explicit by equating real and imaginary parts:

$$\langle x(t)\rangle = \langle x_0\rangle \cos\omega t + \left(\langle p_0\rangle / m\omega\right)\sin\omega t,$$

$$\langle p(t)\rangle = \langle p_0\rangle \cos\omega t - m\omega\langle x_0\rangle \sin\omega t.$$

So we've found Schrodinger's 'best possible' quantum description of a classical oscillator.

Coherent States and the Classical Limit

The phase space of a *classical* simple harmonic oscillator is conveniently parameterized using the complex variable $z = \left(m\omega x + ip\right)/\sqrt{2\hbar m\omega}$. If the oscillator is initially at z_0, it will describe a circle in the complex plane $z(t) = z_0 e^{-i\omega t}$.

If a *quantum* simple harmonic oscillator is initially in a minimum uncertainty state having the same spatial width as the oscillator ground state, and the state has expectation values of position and momentum denoted by $\left(m\omega\langle x_0\rangle, \langle p_0\rangle\right)$, then it is an eigenstate of the annihilation operator $\hat{a}$ with eigenvalue

$$z_0 = \left(m\omega\langle x_0\rangle + i\langle p_0\rangle\right)/\sqrt{2\hbar m\omega},$$

and we label it $|z_0\rangle$,

$$\hat{a}|z_0\rangle = z_0|z_0\rangle.$$

We have established that as this state develops in time, the wave function continues to be a minimal uncertainty wave packet having constant spatial width, and hence continues to be an eigenstate of the annihilation operator, and that the centre of the wave packet in phase space follows *exactly* the path of the classical oscillator,

$$\hat{a}\left|z_0\right\rangle = z\left|z\right\rangle, \quad z = z_0 e^{i\omega t}.$$

A Remark on Notation

We have chosen to work with the original position and momentum variables, and the complex parameter expressed as a function of those variables, throughout. We could have used the dimensionless variables introduced in the lecture on the simple harmonic oscillator,

$$\xi = x/b = x\sqrt{m\omega/\hbar}, \pi = bh/\hbar = p/\sqrt{\hbar m\omega}, \hat{a} = \left(\hat{\xi} + i\hat{\pi}\right)/\sqrt{2}.$$

This would of course also give $z = (\xi + i\pi)\sqrt{2}$ a more compact representation, but one more thing to remember.

It's also common to denote the eigenstates of $\hat{a}$ by $\alpha, \hat{a}\left|\alpha\right\rangle = \alpha\left|\alpha\right\rangle$, very elegant, but we've used z to keep reminding ourselves that this eigenvalue, unlike most of those encountered in quantum mechanics, is a complex number. Finally, some use the dimensionless variables $X = \sqrt{2\hbar/m\omega}x,\ p = \sqrt{1/(2m\omega\hbar)}p,$ differing from ξ, π by a factor of $\sqrt{2}$. The eigenvalue equation for the annihilation operator is very neat in this notation: $\hat{a}\left|z\right\rangle = (X = iP)\left|z\right\rangle$.

The Translation Operator

It's worth repeating the exercise for the simple case of the oscillator initially at rest a distance $\left\langle x_0\right\rangle$ from the centre. This gives a neat tie-in with the *translation* operator.

Let us then take the initial state to be

$$\psi(x,0) = Ce^{-m\omega}\left(x - \left\langle x_0\right\rangle\right)^2/2k = \psi_0\left(x - \left\langle x_0\right\rangle\right)$$

where $\psi_0(x)$ is the ground state wave function so we've moved the packet to the right by $\left\langle x_0\right\rangle$.

Now do a Taylor series expansion (taking $\langle x_0 \rangle$ to be the variable!):

$$\psi\left(x-\langle x_0\rangle\right)=\psi_0\left(x_0\right)-\langle x_0\rangle\frac{d}{dx}\psi\left(x\right)+\frac{\langle x_0\rangle^2}{2!}\frac{d^2}{dx^2}\psi_0\left(x\right)-\ldots$$

$$=e^{-\langle x_0\rangle\frac{d}{dx}}\psi_0\left(x\right).$$

It's clear from this that the *translation operator* $e^{-\langle x_0\rangle\frac{d}{dx}}$ shifts the wave function a distance $\langle x_0 \rangle$ to the right.

Since $\hat{p}=-i\hbar d/dx$, the translation operator can also be written as $e^{-i\langle x_0\rangle\hat{p}/k}$, and from this it can be expressed in terms of $\hat{a},\hat{a}^\dagger$, since

$$\hat{a}=\frac{1}{\sqrt{2\hbar m\omega}}\left(m\omega\hat{x}+i\hat{p}\right),\ \hat{a}^\dagger=\frac{1}{\sqrt{2\hbar m\omega}}\left(m\omega\hat{x}+i\hat{p}\right),$$

($\hat{p},\hat{x}$ being Hermitian) so.

$$\hat{p}=i\sqrt{\frac{\hbar m\omega}{2}}\left(\hat{a}^\dagger-\hat{a}\right).$$

Therefore the displaced ground state wave function can be written

$$\psi_0\left(x-\langle x_0\rangle\right)=e^{-i\langle x_0\rangle\hat{p}/k}\psi_0\left(x\right)$$

$$=e^{\langle x_0\rangle\sqrt{m\omega/2k}\left(\hat{a}^\dagger-\hat{a}\right)}\psi_0\left(x\right)$$

$$=e^{z_0\left(\hat{a}^\dagger-\hat{a}\right)}\psi_0\left(x\right)$$

for *real* $z_0=\langle x_0\rangle\sqrt{m\omega/2\hbar}$, since $\langle p_0 \rangle$ is zero for this initial state (the wave function is real).

In the ket notation, we have established that the minimal uncertainty state centered at x_0, and having zero expectation value for the momentum, is

$$\left|\langle x_0\rangle,0\right|=e^{z_0\left(\hat{a}^\dagger-\hat{a}\right)}\left|0,0\right\rangle.$$

But it's not exactly obvious that this is an eigenstate of $\hat{a}$ with eigenvalue z_0! (As it must be.)

It's worth seeing how to prove that just from the properties of the operators but to do that, we need a couple of theorems concerning exponentials of operators.

First, if the commutator $[A,B]$ commutes with A and B, then $a^{A+B} = e^A e^B e^{-\frac{1}{2}[A,B]}$. This result simplifies the right hand side of the above equation, for

$$e^{za\left(\hat{a}^{\uparrow}-\hat{a}\right)}\left|0,0\right\rangle = e^{z_0\hat{a}^{\uparrow}} e^{-z_0\hat{a}} e^{-za^2[\hat{a}^{\uparrow},\hat{a}^{\uparrow}]/2}\left|0,0\right\rangle$$

$$= e^{-z_0^2/2} e^{z_0\hat{a}^{\uparrow}}\left|0,0\right\rangle$$

where we have used $e^{-z_0 a}\left|0,0\right\rangle = \left|0,0\right\rangle$.

This is simpler, but it's still not obvious that we have an eigenstate of $\hat{a}$: we need the commutator $\left[\hat{a}, e^{z_0\hat{a}^{\uparrow}}\right]$.

The second theorem we need is: if the commutator of two operators $\left[A,B\right] = c$ itself commutes with A and B, then

$$\left[A, e^{\lambda B}\right] = \lambda c e^{\lambda B}.$$

This is easily proved by expanding the exponential.

Applying this to our case,

$$\left[\hat{a}, e^{z_0\hat{a}^{\uparrow}}\right] = z_0 e^{z_0\hat{a}^{\uparrow}}$$

It follows immediately that $e^{-z_0^2/2} e^{z_0\hat{a}^{\uparrow}}\left|0,0\right\rangle$ is indeed an eigenstate of $\hat{a}$ with eigenvalue $z_0 = \left\langle x_0\right\rangle\sqrt{m\omega/2\hbar}$.

(It must also be correctly normalized because the translation

$$\left\langle x_0\right\rangle,0\rangle = e^{z_0\left(\hat{a}^{\uparrow}-\hat{a}\right)}\left|0,0\right\rangle$$

is a unitary operation for real z_0.)

How do we generalize this translation operator to an arbitrary state, with nonzero $\left\langle x\right\rangle,\left\langle p\right\rangle$? Thinking in terms of the complex parameter space *z*, we need to be able to move in both the *x* and the *p* directions, using both $\hat{p} = -i\hbar d/dx$ and $\hat{x} = -i\hbar d/dp$. This is slightly tricky since these operators do not commute, but their commutator is just a number, so this will only affect the overall normalization.

Furthermore, both $\hat{p}$ and $\hat{x}$ are combinations of $\hat{a}, \hat{a}^{\dagger}$, so for the generalization of $e^{-i\langle x_0\rangle \hat{p}/k}$ from real $\langle x_0 \rangle$ to complex *z* to be unitary, it must have an *antihermitian* combination of $\hat{a}, \hat{a}^{\dagger}$ in the exponent a unitary operator has the form $U = e^{iH}$, where *H* is Hermitian, so *iH* is antihermitian.

We are led to the conclusion that

$$|\langle p\rangle,\langle x\rangle\rangle = e^{\left(z\hat{a}^{\dagger} - z^{*}\hat{a}\right)}|0\rangle = |z\rangle,$$

conveniently labeling the coherent state using the complex parameter *z* of its centre in phase space. Since this generalized translation operator is unitary, the new state is automatically correctly normalized.

The Energy Eigenstates

The equation above suggests the possibility of representing the displaced state $|z\rangle$ in the standard energy basis $|n\rangle$. We can simplify with the same trick used for the spatial displacement case in the last section, that is, the theorem $e^{A+B} = e^{A}e^{B}e^{-\frac{1}{2}[A,B]}$ where now $A = z\hat{a}^{\dagger}$, $B = -z^{*}\hat{a}$:

$$|z\rangle = e^{\hat{z}^{\dagger} - z^{*}\hat{a}}|0\rangle = e^{-|z|^2/2}e^{z\hat{a}^{\dagger}}e^{-z^{*}\hat{a}}|0\rangle = e^{-|z|^2/2}e^{z\hat{a}^{\dagger}}|0\rangle$$

using $e^{-z^{*}\hat{a}}|0\rangle = |0\rangle$ since $\hat{a}|0\rangle = 0$.

It is now straightforward to expand the exponential:

$$|z\rangle = e^{-|z|^2/2}e^{z\hat{a}^{\dagger}}|0\rangle = e^{-|z|^2/2}\left(1 + za^{\dagger} + \frac{\left(za^{\dagger}\right)^2}{2!} + \ldots\right)|0\rangle$$

and recalling that the normalized energy eigenstates are

$$|n\rangle = \frac{\left(a^\dagger\right)^n}{\sqrt{n!}}|0\rangle$$

we find

$$|z\rangle = e^{-|z|^2/2}\left(|0\rangle + z + |1\rangle + \frac{z^2}{\sqrt{2!}}|2\rangle + \frac{z^3}{\sqrt{3!}}|3\rangle + ...\right).$$

Time Development of an Eigenstate of a *Using the Energy Basis*

Now that we have expressed the eigenstate $|z\rangle$ as a sum over the eigenstates $|n\rangle$ of the Hamiltonian, finding its time development in this representation is straightforward.

Since $|n(t)\rangle = e^{-in\omega t}|n\rangle$,

$$|z(t)\rangle = e^{-|z_a|^2/2}\left(|0\rangle + z_0 e^{-i\omega t}|1\rangle + \frac{z_0^2 e^{-2i\omega t}}{\sqrt{2!}}|2\rangle + \frac{z_0^3 e^{-3i\omega t}}{\sqrt{3!}}|3\rangle + ...\right)$$

which can be written

$$|z(t)\rangle = e^{-|z_0|^2/2} e^{z_0 e^{-m}\hat{a}^\dagger}|0\rangle,$$

equivalent to the result derived earlier.

Some Properties of the Set of Eigenstates of a

In quantum mechanics, any physical variable is represented by a Hermitian operator. The eigenvalues are real, the eigenstates are orthogonal (or can be chosen to be so for degenerate states) and the eigenstates for a complete set, spanning the space, so any vector in the space can be represented in a unique way as a sum over these states.

The operator $\hat{a}$ is not Hermitian. Its eigenvalues are *all the numbers in the complex plane*. The eigenstates belonging to different eigenvalues are never orthogonal, as is immediately obvious on considering the ground state and a displaced ground state. The overlap does of course decrease rapidly for states far away in phase space.

The state overlap can be computed using

$$|z(t)\rangle = e^{-|z_0|^2/2} e^{z\hat{a}^\dagger}|0\rangle:$$

$$\langle w|z\rangle = \left\langle 0\left|e^{w^*\hat{a}}e^{-|w|^2/2}e^{-|z|^2/2}e^{z\hat{a}^\dagger}\right|0\right\rangle$$

and we can then switch the $e^{-w^*\hat{a}}, e^{z\hat{a}^\dagger}$ operators using the theorem $e^B e^A = e^A e^B e^{-[A,B]}$, then since we're left with

$$\langle w|z\rangle = \left\langle 0\,|\,e^{w^*z}e^{-|w|^2/2}e^{-|z|^2/2}\,|\,0\right\rangle,$$

from which $|\langle w|z\rangle|^2 = e^{-|w-z|^2}$.

Finally, using

$$|z\rangle = e^{-|z|^2/2}\left(|0\rangle + z + |1\rangle + \frac{z^2}{\sqrt{2!}}|2\rangle + \frac{z^3}{\sqrt{3!}}|3\rangle + \ldots\right),$$

we can construct a unit operator using the $|z\rangle$,

$$1 = \iint \frac{dxdy}{\pi}|z\rangle\langle z|$$

where the integral is over the whole complex plane $z = x + iy$ (this x is not, of course, the original position x, recall for the wave function just displaced along the axis $z_0 = \langle x_0\rangle\sqrt{m\omega/2\hbar}$). Therefore, the $|z\rangle$ span the whole space.

Some Exponential Operator Algebra

Suppose that the commutator of two operators A, B

$$[A,B] = c,$$

where c commutes with A and B, usually it's just a number, for instance 1 or $i\hbar$.

Then

$$\left[A,e^{\lambda B}\right] = \left[A, 1 + \lambda B + \left(\lambda^2/2!\right)B^2 + \left(\lambda^3/3!\right)B^3 + \ldots\right]$$

$$= \lambda c + \left(\lambda^2/2!\right)2Bc + \left(\lambda^3/3\right)3B^2c + \ldots$$

$$= \lambda c e^{\lambda B}.$$

That is to say, the commutator of A with $e^{\lambda B}$ is proportional to $e^{\lambda B}$ itself.

That is reminiscent of the simple harmonic oscillator commutation relation $\left[H,a^{\dagger}\right]=\hbar\omega a^{\dagger}$ which led directly to the ladder of eigenvalues of H separated by $\hbar\omega$. Will there be a similar 'ladder' of eigenstates of A in general?

Assuming A (which is a general operator) has an eigenstate $|a\rangle$ with eigenvalue a,

$$A|a\rangle = a|a\rangle$$

Applying $\left[A,e^{\lambda B}\right]=\lambda ce^{\lambda B}$ to the eigenstate $|a\rangle$:

Therefore, unless it is identically zero $e^{\lambda B}|a\rangle$, is *also* an eigenstate of A, with eigenvalue $a+\lambda c$. We conclude that instead of a *ladder* of eigenstates, we can apparently generate a whole *continuum* of eigenstates, since λ can be set arbitrarily.

To find more operator identities, premultiply $\left[A,e^{\lambda B}\right]=\lambda ce^{\lambda B}$ by $e^{-\lambda B}$ to find:

$$\begin{aligned} e^{-\lambda B}AB^{\lambda B} &= A+\lambda[A,B] \\ &= A+\lambda c. \end{aligned}$$

This identity is *only* true for operators A, B whose commutator c is a number. (Well, c *could* be an operator, provided it still commutes with both A and B).

$$e^{A+B}=e^{A}e^{B}e^{-\frac{1}{2}[A,B]}.$$

The proof (due to Glauber, given in Messiah) is as follows:

Take $f(x)=e^{Ax}e^{Bx}$,

$$\begin{aligned} df/dx &= Ae^{Ax}e^{Bx}+Ae^{Ax}e^{Bx}B \\ &= f(x)\left(e^{-BX}Ae^{Bx}+B\right) \\ &= f(x)\left(A+[A,B]+B\right). \end{aligned}$$

It is easy to check that the solution to this first-order differential equation equal to one at $x = 0$ is

$$f(x) = e^{x(A+B)} e^{\frac{1}{2}x^2[A,B]}$$

so taking $x = 1$ gives the required identity, $e^{A+B} = e^A e^B e^{-\frac{1}{2}[A,B]}$.

It also follows that $e^B e^A = e^A e^B e^{-[A,B]}$ provided as always that $[A, B]$ commutes with A and B.

9

The Hydrogen Atom

Energy and the Hydrogen Atom

In 1885 a Swiss secondary school teacher named Johann Jacob Balmer published a short note (entitled "Note on the Spectral Lines of Hydrogen", *Annalen der Physik und Chemie* 25, 80-5) in which he described an empirical formula for the four most prominent wavelengths of light emitted by hydrogen gas. These wavelengths had been measured with great precision by Vogel and Huggins, giving the four values 6562.10, 4860.74, 4340.10, and 4101.20 Angstroms (10^{-10} m). Balmer's note does not make clear whether he was also aware of the measured series limit, $\lambda_\infty = 3645.6$ A, or whether he deduced this himself. In any case, one can find by numerical experimentation that the four characteristic wavelengths are closely proportional to the following products of small primes.

$1512 = 2^3 \times 3^3 \text{ X } 3^3 \times 7 \qquad 1120 = 2^5 \times 5 \times 7$

$1000 = 2^3 \times 5^3 \qquad\qquad\qquad 945 = 3^3 \times 5 \times 7$

Three of these are divisible by 2^3, three are divisible by 5, three are divisible by 7, and two are divisible by 3^3. Thus we can easily express these numbers as simple fractional multiples of $840 = 2^3 \times 3 \times 5 \times 7$, which corresponds to the series limit $\lambda_\infty = 3645.6$ A. It may have been just this kind of numerical experimentation that led Balmer to recognize that the four prominent wavelengths are given very closely by $(9/5)\lambda_\infty$, $(16/12)\lambda_\infty$, $(25/21)\lambda_\infty$, and $(36/32)\lambda_\infty$. He also noticed that the numerators of the coefficients are consecutive squares, and each denominator is 4 less than the numerator. He speculated that the pattern would continue up to the series limit, which is indeed the case. In terms of the wave number κ (=1/λ), Balmer's formula can be written as

$$\kappa = \frac{n^2}{\lambda_\infty}\left(\frac{j^2 - n^2}{n^2 j^2}\right) = R_H\left(\frac{1}{n^2} - \frac{1}{j^2}\right) \quad n = 2;\ j = 3,4,5...$$

where $R = n^2/\lambda_\infty$ with $n = 2$. The parameter R_H is now called Rydberg's constant for hydrogen, and the best empirical value is 10967757.6 m^{-1}. As Balmer also speculated, if we take different values of n we get different series of spectral lines. The series with n = 1, 2, 3, 4, and 5 are now known as the Lyman, Balmer, Paschen, Brackett, and Pfund series, respectively, which characteristic the spectral lines of the hydrogen atom. (The Balmer series was observed first because its frequencies are in the visible and near ultra-violet range.) This is an outstanding example of a successful empirical fit (like Bode's Law in astronomy) for a class of physical phenomena that was not based on any underlying physical model or theory, i.e., no reason was known for why wavelengths of light emitted from a hydrogen atom should exhibit this pattern.

In classical terms a hydrogen atom consists of a proton and an electron bound together by their mutual electrical attraction. To keep them from collapsing together, we might imagine that the electron is revolving in 'orbit' around the proton, similar to a planet revolving around the Sun, with the centrifugal force balancing the electrical attraction. However, this model is not satisfactory, because the electron would be continuously accelerating, and according to classical theory an accelerating charge radiates energy in the form of electro-magnetic waves. As a result, the orbiting electron would very quickly radiate away all of its kinetic energy and spiral into the proton. Thus the existence of stable atoms was inexplicable in the context of classical physics, as was the characteristic set of discrete energy levels of atoms.

By the early 1900s it had become clear that classical electrodynamics was inadequate to account for the behaviour of either the electromagnetic field or of elementary particles. In 1900 Max Planck had shown in his study of black-body radiation that it is necessary to quantize the energy of electromagnetism in order to avoid the 'untra-violet catastrophe', and he introduced the fundamental constant h. In 1905 Einstein made the even more radical proposal that in some respects electromagnetic wave energy propagates as if it consists of small packets (photons) with many of the characteristics of particles, each photon having an energy E related to the wave frequency ν by $E = h\nu$.

In 1913 Niels Bohr developed a new representation of the hydrogen atom by combining classical ideas with a few additional postulates that were suggested by the nascent quantum concepts of Planck and Einstein. First, he assumed that the angular momentum of an electron in orbit around the nucleus must be an integer multiple of $\hbar$ (Planck's constant h divided by 2π). It follows that only a certain set of discrete energy levels may occur. Second, he assumed that an electron radiates energy only when it makes a transition from one stable orbit to another of lower energy. If ΔE is the difference in energy levels, then he assumed that the transition resulted in the emission of a photon with this amount of energy, and hence with the frequency $\nu = \Delta E/h$ in accord with Einstein's postulate. Armed with these postulates, Bohr reasoned that an electron of mass m orbiting a proton (of much greater mass) at radius r would satisfy the classical force balance

$$\frac{e^2}{4\pi\varepsilon_0 r^2} = m\frac{v^2}{r}$$

and the total energy (kinetic plus potential) has the classical value

$$E = \frac{1}{2}mv^2 + (-e)V$$

$$= \frac{e^2}{8\pi\varepsilon_0 r} - \frac{e^2}{4\pi\varepsilon_0 r} - \frac{e^2}{8\pi\varepsilon_0 r}.$$

Likewise the angular momentum has the classical value

$$L = mvr = \sqrt{\frac{me^2 r}{4\pi\varepsilon_0}}$$

Bohr then imposed his quantization assumption, asserting that L must be an integer multiple of $\hbar$ Setting $L = n\hbar$ and solving the above equation for r, we get

$$r = \frac{4\pi\varepsilon_0 n^2\hbar^2}{me^2}.$$

Substituting into the expression for the energy E gives the corresponding quantized energy levels of the hydrogen atom

$$E_n = -\frac{me^4}{2\left(4\pi\varepsilon_0\right)^2 \hbar^2 n^2}$$

This provides a nice rationale for Balmer's empirical formula, because it implies that the frequency of the emitted light when an electron makes a transition from the jth to the nth energy level is

$$n = \frac{\Delta E}{h} = \frac{me^4}{2h\left(4\pi\varepsilon_0\right)^2 \hbar^2}\left(\frac{1}{n^2} - \frac{1}{j^2}\right)$$

Since $\lambda \nu = c$ we have $\kappa = \nu/c$ and therefore

$$k = \left(\frac{me^4}{h^3 \varepsilon_0^2 c}\right)\left(\frac{1}{n^2} - \frac{1}{j^2}\right).$$

Actually, to be more accurate, the mass m in this expression should be replaced with the "reduced mass" $mM/(m+M)$ where M is the mass of the proton (or, more generally, the nucleus), just as in classical orbital mechanics. Then the coefficient in the above expression is identified with Rydberg's constant R_H for the hydrogen atom, and using the values of the fundamental constants

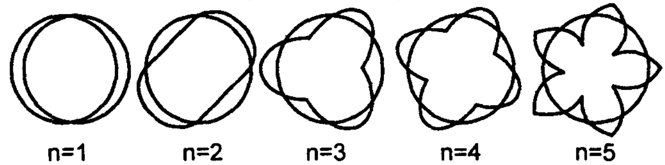
c = 2.99792458 X 10^8 m/s
h = 6.626176 X 10^{-34} j X s
e_0 = 8.854187818 X 10^{-12} F/m
e = 1.6021892 X 10^{-19} C
m = 9.109534 X 10^{-31} kg
m = 1.67264 X 10^{-27} kg

we can compute

$$R_H = \left(\frac{mM}{m+M}\right)\frac{e^4}{h^3 \varepsilon_0^2 c} = 10967757.78 \text{ m}^{-1}.$$

This is in remarkable agreement with the measured value of 10967757.6 ± 1.2 m^{-1} from spectroscopic data. Nevertheless, Bohr's model of the atom is not completely satisfactory, partly because of the ad hoc nature of its premises, and also because it's representation of electrons as tiny particles with definite trajectories is not viable in a wider context.

A somewhat plausible justification for Bohr's quantization postulate came in 1924 when Louis de Broglie developed the idea that particles of matter on the smallest scale exhibit wave-like properties, complementing Einstein's suggestion that electromagnetic waves exhibit particle-like properties. The de Broglie wavelength for the matter wave corresponding to a particle with momentum p is $\lambda = h/p$, and if we stipulate that the circumference $2\pi r$ of a circular orbit of radius r must be an integer multiple of the wavelength, we have $2\pi r/\lambda = 2\pi r p/h = n$ for some positive integer n. Since the angular momentum is $L = pr$, this immediately gives Bohr's quantization postulate $L = n\hbar$. This 'orbital wave quantization' is illustrated for $n = 1$ through 5 in the figures below.



However, despite the plausibility of this approach, Bohr's model of the hydrogen atom, even with de Broglie's justification and with subsequent refinements by Sommerfeld, is now considered obsolete, having been superceded by a more thorough-going wave mechanics developed by Erwin Schrodinger in 1925. (This new theory was subsequently shown to be essentially identical to the 'matrix mechanics' already developed by Werner Heisenberg in 1924.)

Schrodinger's wave mechanics postulates that a particle is characterized by a complex-valued wave function Ψ(x,y,z,t) whose squared norm at any point equals the probability density for the particle to be found at that point. (The probability interpretation of Schrodinger's wave function was first proposed by Max Born.) In addition, Schrodinger postulated that, in a region where there is a potential field $V(x,y,z,t)$, the wave function ψ of a particle is governed by the equation

$$-\frac{\hbar^2}{2m}\left[\frac{\partial^2\psi}{\partial x^2}+\frac{\partial^2\psi}{\partial y^2}+\frac{\partial^2\psi}{\partial z^2}\right]+V\psi$$

$$= i\hbar\frac{\partial\psi}{\partial t}$$

It's possible to give a plausibility argument for this equation, but here we will just take it as given. Expressing the spatial Laplacian (the quantity in the square brackets) in terms of polar coordinates, and considering just the radial part, this equation is

$$-\frac{\hbar^2}{2m}\left[\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial\psi(r,t)}{\partial r}\right)\right]+V(r,t)\psi(r,t)=i\hbar\frac{\partial\psi(r,t)}{\partial t}$$

Furthermore, if the potential field V does not change with time, we can separate the variables by expressing $\Psi(r, t)$ as a product of a spatial part and a temporal part, i.e., we have functions $\varphi(r)$ and $\omega(t)$ such that $\Psi(r, t) = \varphi(\mathrm{r})\omega(\mathrm{t})$. Substituting into the above equation and dividing through by $\Psi(r,t)$ gives

$$-\frac{\hbar^2}{2m\varphi(r)}\left[\frac{1}{r^2}\frac{d}{dr}\left(r^2\frac{d\varphi(r)}{dr}\right)\right]+V(r)=\frac{i\hbar}{\omega(t)}\frac{d\omega}{dt}$$

Since r and t are independent, the left and right hand sides of this equation must both equal a constant, which we will call E, so the time-independent radial Schrodinger equation in this simple case is

$$-\frac{\hbar^2}{2m}\left[\frac{1}{r^2}\frac{d}{dr}\left(r^2\frac{d\varphi(r)}{dr}\right)\right]+V(r)\varphi(r)=E\varphi(r)$$

Now, as mentioned previously, for the region around a charged proton the potential energy due to the Coulomb force is given by

$$V(r)=-\frac{e^2}{4\pi\varepsilon_0 r}$$

Inserting this into the Schrodinger equation, evaluating the nested derivatives, and re-arranging terms, we get

$$\frac{d^2\varphi}{dr^2}+\frac{2}{r}\frac{d\varphi}{dr}+\frac{2m}{\hbar^2}\left[\frac{e^2}{4\pi\varepsilon_0 r}+E\right]\varphi=0 \quad (1)$$

For sufficiently large values of r the terms with r in the denominator will be negligible, so the equation will reduce to

$$\frac{d^2\varphi}{dr^2}+\frac{2m}{\hbar^2}E\varphi=0$$

which has the solution

$$\varphi(\mathrm{r})=\mathrm{e}^{-\mathrm{cr}}$$

$$\text{Where C} = C=\sqrt{-\frac{2mE}{h^2}}.$$

To exploit this asymptotic result, and without loss of generality, we can consider a general solution of the form $\varphi(r) = F(r)e^{-Cr}$. Substituting into the Schrodinger equation, evaluating the derivatives, and dividing through by e^{-Cr}, we get

$$\frac{d^2F}{dr^2}+2\left(\frac{1}{r}-C\right)\frac{dF}{dr}+\frac{2}{r}\left(B-C\right)F=0$$

where

$$B=\frac{me^2}{4\pi\varepsilon_0 h^2}$$

If the function $F(r)$ is analytic it can be represented by a power series, i.e., it can be expressed in the form

$$F(r) = f_0 + f_1 r + f_2 r^2 + \ldots$$

Substituting into the equation for F, collecting terms by powers of r, and setting the coefficients of these terms to zero, we arrive at the conditions

$$f_k = 2\frac{kC-B}{k\left(k+1\right)}f_{k-1} \quad \text{for k} = 1, 2, \ldots.$$

For sufficiently large k these expressions approach $f_k = [2C/(k+1)]f_{k-1}$, which is the series for e^{2Cr}, and hence $F(r)e^{-Cr}$ goes to e^{Cr}, which increases to infinity as r increases. Therefore, in order to give a solution that goes to zero as r goes to infinity, we must impose the requirement that the series for $F(r)$ terminates after a finite number of terms. This occurs if and only if $nC = B$ for some positive integer n. Hence the necessary and sufficient condition for the solution to approach zero as r increases is

$$n\sqrt{-\frac{2mE}{h^2}}=\frac{me^2}{4\pi\varepsilon_0 h^2}$$

Squaring both sides and solving for E gives the allowable energy levels

$$E_n=-\frac{me^4}{(4\pi\varepsilon_0)^2 2\hbar^2 n^2}.$$

Which is identical to the discrete energy levels of Bohr's model discussed previously. It's worth noting that the quantization of energy levels here is not the result of quantized angular momentum or orbital standing-waves. It arises from an analysis of the purely radial component of the Schrodinger wave equation of the ground state, which is spherically symmetrical and has an angular quantum number of zero.

Superficially it isn't obvious that the 'realistic solutions' of must be quantized, so it's worthwhile to examine the solution technique more closely to understand clearly how the quantization arises. First, notice that if we had tried to find a series solution of directly by inserting a series $\varphi(r) = \varphi_0 + \varphi_1 r + \varphi_2 r^2 + \ldots$ we would have arrived at a set of conditions involving three consecutive coefficients. This can be seen by inspection, because when we carry out the differentiations and collect the coefficients of r^k, any term of the original differential equation of the form $r^s\, d^q\varphi/dr^q$ contributes a quantity involving φ_{k+q-s}. Hence the four terms of contribute quantities involving φ_{k+2}, φ_{k+2}, φ_{k+1}, and φ_k respectively. In contrast, the four terms of contribute quantities involving f_{k+2}, f_{k+2}, f_{k+1}, and f_{k+1} respectively, so the power series conditions enable us to determine each f_{k+2} as a multiple of f_{k+1}. We originally motivated the solution form $F(r)e^{-cr}$ based on the asymptotic solution for large r, but we could also have justified it based on the fact that this transformation leads to a differential equation whose power series solution is subjected to conditions on just two consecutive coefficients. The fact that such a transformation exists is crucial for the quantization.

This leads us to consider the general conditions in which such a transformation exists. Suppose we have a second-order differential equation of the form

$$\alpha(x)\frac{d^2y}{dx^2} + \beta(x)\frac{dy}{dx} + \gamma(x)y = 0$$

where α, β, and γ are rational functions of x. Since we can multiply through by arbitrary polynomials in x, we can assume without loss of generality that α, β, γ are polynomials in x. If we postulate a solution of the form $y(x) = f(x)e^{g(x)}$ and substitute into this equation, we get

$$[\alpha]\ddot{f} + [2\alpha\ddot{g} + \beta]\ddot{f} + [\alpha\ddot{g} + \alpha\dot{g} + \beta\dot{g} + \gamma]f = 0$$

In order for the series solution for $f(x)$ to have conditions on just the sets of two consecutive coefficients, there must be an integer d such that the coefficient of f'' contains only terms in x^{d+1} and x^{d+2}, and the coefficient of f' contains only terms in x^d and x^{d+1}, and the coefficient of f contains only terms in x^d. We seek a function $g(x)$ such that these conditions are satisfied. In the case of equation we have (after multiplying through by r) an equation of the form with $\alpha(x) = x$, $\beta(x) = 2$, and $\gamma(x) = 2B - C^2x$ where B and C are the constants defined previously. Therefore, from the condition on α, we see that d is either 0 or 1, so we need a function $g(x)$ such that $2xg' + 2$ involves only terms in x^0 and x^1, or only terms in x^1 and x^2. Since it certainly involves a term in x^0, the remaining term must be in x^1, so $g'(x)$ must be a constant. Also, the coefficient of f must be in x^0 (i.e., a constant), so we have $x(g')^2 + 2g' + 2B - C^2x = K$. Therefore we must have $(g')^2 = (C)^2$, which leads to the transformation $y(x) = f(x)e^{-Cx}$ as expected.

The superiority of the wave mechanical model of the hydrogen atom over Bohr's model is immense, because it not only duplicates and (in a sense) "explains" the quantized energy levels, it actually gives the complete probability density functions for the various possible stationary states. Using the recursive formula, we can evaluate the coefficients of the polynomial $F(r)$ for each value of the quantum number n. Combining these polynomials with the exponential parts, we have the wave functions of the first few states.

$$\varphi_1(r) = f_0\, e^{-Cr}$$

$$\varphi_2(r) = (1 - Cr)f_0\, e^{-Cr}$$

$$\varphi_3(r) = \left(1 - 2Cr + \frac{2}{3}C^2r^2\right)f_0\, e^{-cr}$$

$$\varphi_4(r) = \left(1 - 3Cr + 2C^2r^2 - \frac{1}{3}C^3r^3\right)f_0\, e^{-cr}$$

$$\varphi_5(r) = \left(1 - 4Cr + 4C^2r^2 - \frac{4}{3}C^3r^3\right)f_0\, e^{-Cr}$$

Recall that $B = nC$ and B is composed entirely of fundamental constants, independent of n, and it has units of inverse length. Letting a_0 denote the length $1/B = 1/(nC)$, and using m instead of the more accurate reduced mass, we have

$$a_0 = \frac{4\pi\varepsilon_0\hbar^2}{me^2}$$

and we can substitute $1/(na_0)$ for C in the preceding wave functions and clear fractions. In terms of the normalized radius parameter $\rho = r/a_0$ the wave functions are

$$\varphi_1(\rho) = f_0\, e^{-\rho}$$

$$\varphi_2(\rho) = (2-\rho)\frac{f_0}{2}\, e^{-\rho/2}$$

$$\varphi_3(\rho) = (27-18\rho + 2\rho^2)\frac{f_0}{27}\, e^{-\rho/3}$$

$$\varphi_4(\rho) = (192 - 144\rho + 24\rho^2 - \rho^3)\frac{f_0}{192}\, e^{-\rho/4}$$

$$\varphi_5(\rho) = (9375 - 7500\rho + 1500\rho^2 - 100\rho^3 + 2\rho^4)\frac{f_0}{9375}\, e^{-\rho/5}$$

Each of these wave functions includes a constant factor f_0. To determine the value of this factor, recall that the squared norm of the wave function is the probability density, and so the integral of this quantity over all of space must equal 1. The volume of an incremental spherical shell of radius r and thickness dr is $4\pi r^2 dr$ so the probability integral is

$$\int_0^\infty 4\pi r^2 \left(\varphi_n(r)\right)^2 dr = 4\pi a_0^{\,3} \int_0^\infty \rho^2 \left(\varphi_n(\rho)\right)^2 dp = 1$$

For example, to find the constant factor f_0 for the case $n = 1$ we insert the wave function into this equation and evaluate the integral to give

$$4\pi a_0^{\,3} f_0^{\,2} \int_0^\infty \rho^2 \left(e^{-\rho}\right)^2 dp = \pi a_0^{\,3} f_0^{\,2} = 1$$

Therefore we have $f_0 = (\pi a_0^{\,3})^{-1/2}$, and the complete wave function for the ground state of the hydrogen atom is

$$\varphi_1(r) = \frac{e^{-r/a_0}}{\sqrt{\pi a_0^{\,3}}}$$

In accord with our choice of nomalizing factors, the probability density for finding the electron in an incremental shell of radius r is $\delta_n(r) = 4\pi r^2 \varphi_{-n}(r)^2$. This is plotted in the figure below for the first few values of n.

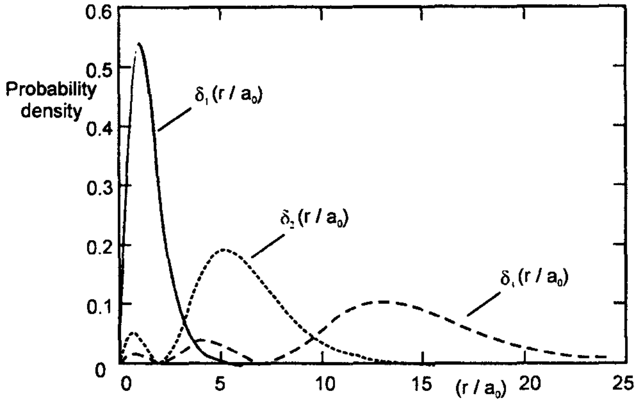


Fig. 1 Probability Density of Finding Electrons.

Throughout this discussion we've ignored the angular components of the electron's wave function, 'effectively assuming that it has zero angular momentum, so the only non-zero quantum number was the radial one. This was all based on taking just the radial part of the Laplacian in the Schrodinger equation. If we had taken the angular parts we would have found that those too are associated with quantum numbers 0,1,2,..., and they contribute to the overall orbital wave function. However, the essential features of Schrodinger's approach to the hydrogen atom are already apparent in the purely radial part.

The Wave Equation and Permutation of Rays

The usual wave equation in one space and one time dimension is

$$\frac{\partial^2 \Psi}{\partial x^2} = \frac{1}{c^2} \frac{\partial^2 \Psi}{\partial t^2}$$

This same equation applies to spherically symmetrical waves in three-dimensional space if we replace x with the radial distance r from the centre of the disturbance, and if we replace the wave function ψ with $\phi = r\psi$. Therefore, the solution of the above equation is relevant to many important physical phenomena. The general analytical solution is not difficult to find, but we can gain useful insights into wave propagation if we consider this equation in the form of finite differences.

For any point x, t in the medium we can consider the four neighbouring points at incremental distances in space and time.

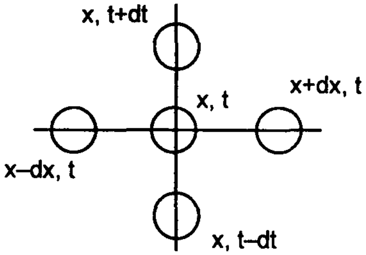


Fig. 2 Distance in Space and time

Expressing the second partial derivative of the wave function with respect to x in terms of finite differences around the central point, we have

$$\frac{\partial^2\psi}{\partial x^2} = \frac{\dfrac{\psi(x+dx,t)-\psi(x+dx+t)}{dx} - \dfrac{\psi(x,t)-\psi(x-dx,t)}{dx}}{dx}$$

$$= \frac{\psi(x+dx,t)-2\psi(x,t)+\psi(x-dx,t)}{(dx^2)}$$

Likewise the second partial derivative of the wave function with respect to t can be represented in the form

$$\frac{\partial^2\psi}{\partial t^2} = \frac{\psi(x,t+dt)-2\psi(x,t)+\psi(x,t-dt)}{(dt^2)}$$

If we then choose our units of space and time so that $dx = c\ dt$, we can substitute these finite difference expressions into the wave equation, simplify, and multiply through by $(dt)^2$ to give

$$\psi(x, t + dt) + \psi(x, t - dt) = \psi(x, dx, t) + \psi(x - dx, t)$$

Notice that the value of the wave function at the central point drops out, so the finite difference equation operates only on the four corners of the surrounding cell. Thus the wave equation simply expresses the requirement that the sum of the values of the wave function just before and just after a given event equals the sum of the values on either side of the event. (The average of its neighbours in time equals the average of its neighbours in space.) With this simple rule, we can examine how a disturbance propagates.

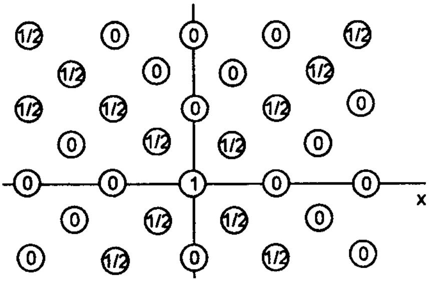


Fig. 3 Grid of such Wave function at the central point.

The values shown in each node of the grid represent the value of the wave function at that point in space and time. A disturbance of magnitude 1 is posited at the origin, with zero specified at every other spatial location at that instant. (Alternately, we can specify zero for every other instant at the origin.) Assuming both spatial and temporal symmetry, the resulting propagation of this disturbance is indicated by the nodes marked with '1/2'. Since the space increment *dx* equals *c* times the time increment, and since the disturbance propagates *at* ± *dx*/*dt*, this shows that the disturbance propagates with the speed *c*. Also, it's clear that solutions of the wave equation are linear, in the sense that the sum of any two solutions is another solution.

It's often most useful to express the basic difference equation in one of the two forms

$$\psi(x, t + dt) - \psi(x + dx, t) = \psi(x - dx, t) - \psi(x, t - dt)$$
$$\psi(x, t + dt) - \psi(x - dx, t) = \psi(x + dx, t) - \psi(x, t - dt)$$

These equations show explicitly that the change in the wave function along one edge of a diamond cell equals the change along the opposite edge .

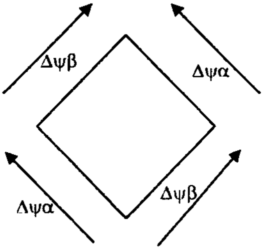


Fig. 4 Diamond Cell

By transitivity with adjoining cells, it immediately follows that the change in the wave function is invariant along opposite edges of any rectangular region oriented orthogonally to these cells. Consequently, the sums of the wave function values on opposite vertices of any such rectangular region are equal.

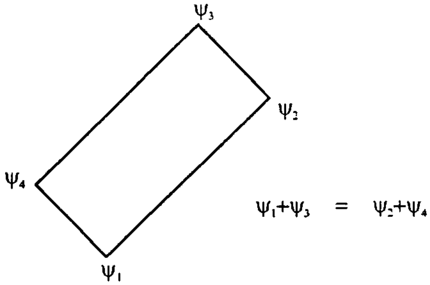


Fig. 5 Rectangular Region

In general we have $\psi(x, t) + \psi(x + c\Delta t_1 + c\Delta t_2, \mathrm{t} + \Delta t_1 + \Delta t_2) = \psi(x + c\Delta t_1, t + \Delta t_2$ for any intervals Δt_1 and Δt_2, *no matter how large*. It's interesting that partial differential equations are often regarded as characteristic of local processes, and the wave equation is the archetype partial differential equation, but we find that the basic relationships can just as well be expressed in non-local form. This illustrates how problematical it is to define a flow of causality for a deterministic process, even when the governing equation can be expressed in the form of a partial differential equation. It also suggests that the 'topology of implication' of the wave function is not Euclidean, but is more accurately represented by the indefinite 'metric' of Voigt and Lorentz, with singular measures along the diagonals.

Of course, as d'Alembert observed, in terms of the coordinates $u = x + t$ and $v = x - t$ the wave equation (with unit c) reduces to

$$\frac{\partial^2 \psi}{\partial u \partial v} = 0$$

which implies that $\partial\psi/\partial u$ is strictly a function of u, and likewise $\partial\psi/\partial v$ is strictly a function of v. Consequently the entire solution can be expressed as the sum of two single-variable functions, $\psi(x, t) = f(x + t) + g(x - t)$. Another way of encoding this solution is to say that to each point x in the one-dimensional space we assign two values, $\partial\psi/\partial u$ and $\partial\psi/\partial v$. The entire spacetime solution is projected (along lines of constant u and lines of constant v) from this single time-slice. Also, notice in particular that if $\psi(u, v)$ is a solution of the wave equation, then so is $\psi(\alpha u, \beta v)$ for any constants α, β.

It's not difficult to show that if $\psi(x, t)$ is a solution of the wave equation in terms of x and t, and if we postulate a linear transformation between x, t and X, T of the form

$$x = AX + BT \qquad t = CX + DT$$

then $\psi(X, T)$ is a solution of the wave equation in terms of X and T if and only if

$$A^2 + C^2 = B^2 + D^2 \qquad AC = BD$$

From this it follows that

$$(A + C^2) = (B + D)^2 \, (A - C)^2 = (B - D)^2$$

and therefore we have $(A + C) = \pm(B + D)$ and $(A - C) = \pm(B - D)$. Consequently the transformation from x,t to X,T can be written in the form

$$x + t = (A + C)(X + T) \; x - t = (A - C)(X - T)$$

which confirms that any re-scaling of the u,v variables preserves the solution.

Although the Voigt-Lorentz transformation is singled out as the continuous linear inevitable transformation that preserves the wave equation, it's obvious from the preceding that the wave equation is actually preserved by a much larger class of transformations. In fact, returning to our finite difference grids, we can see that the wave equation is preserved under any permutation of the constant-u lines, and under any permutation of the constant-v lines (i.e., the lines of constant $x + t$ and of constant $x - t$). In physical terms, the wave equation if preserved under any permutation of the light rays.

With two spatial dimensions and one time dimension we have a pencil of light rays intersecting at each point, forming forward and backward cones. Again the transverse derivatives are constant along light rays, so the entire spacetime solution can be represented by the projection onto a single time slice, where at each point we have a range of angles from 0 to 2π. In effect, this is a three dimensional space where one of the dimensions is finite and curled up cylindrically. A light ray in spacetime maps to a single point in this projected space, whereas a point (event) in spacetime can be regarded as the entire pencil of rays that intersect at that point, which imply that it maps to thread that wraps around the cylindrical dimension of the projected space as illustrated below.

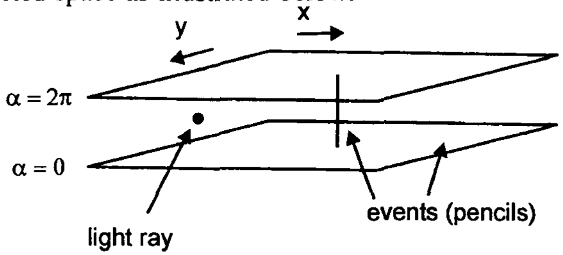


Fig. 6 Pencil of Light Rays Intersecting

With three spatial dimensions and one time dimension the rays of light comprise an expanding shell of light converging on and emanating from each point (event). These rays can be projected onto a single 3D time slice with a closed curled-up dimensional spherical surface at each point. In other words, it can be modeled as $E_3 \times S_2$, the Cartesian product of three-dimensional Euclidean space and two-dimensional spherical surface. Again each individual light ray is a point in this projected space, whereas the pencil of rays intersecting at a given event maps to a closed manifold that wraps around the spherical sub-space. (If at each point in the three-dimensional space we have not only the two-dimensional manifold of spatial directions, but also a spin orientation about each direction, then the full six-dimensional space can be represented by the Cartesian product $E_3 \times S_2 \times S_1$.)

In full 3+1 dimensional spacetime the spherically symmetrical wave equation can be reduced to the same form as the 1+1 dimensional equation, except that the space wave function is normalized by r, and the parameter x is replaced with the radius r, so each 'light ray' actually represents a sequence of expanding and converging shells. The above reasoning shows that we can permute any of these spatially concentric sequences of shells and still preserve the wave equation. Of course, we can also apply a Lorentz transformation and then a permutation, so we can effectively permute any two 'pencils' of light shells that are within each others past or forward light cones. In this sense, we could say that two events (associated with their respective light cones) are causally ordered if and only if they can be permuted while preserving the wave equation.

Does Relativistic Mass Imply Special Relativity?

In a collection of essays on the subject of special relativity Richard Feynman presents the formula for relativistic mass

$$m = \frac{m_0}{\sqrt{1 - v^2 / c^2}}$$

and then remarks that : For those who want to learn just enough about it so they can solve problems, that is all there is to the theory of relativity—it just changes Newton's laws by introducing a correction factor to the mass.

Unfortunately he gives no explicit explanation of this assertion. Later he discusses how the relativistic mass formula can be derived from the Lorentz transformation, but that's the reverse of what's needed to support the above claim, i.e., he needs to show that the Lorentz transformation follows from the relativistic mass formula.

Whenever you see a sweeping statement that a tremendous amount can come from a small number of assumptions, you always find that it is false. There are usually a large number of implied assumptions that are far from obvious if you think about them sufficiently carefully.

Nevertheless, he did claim that all of special relativity follows from, so it's interesting to consider in what sense this claim is valid. Some authors—especially those who disapprove of the notion of 'relativistic mass'—have argued that Feynman was simply wrong, i.e., they assert that special relativity does not follow from relativistic mass.

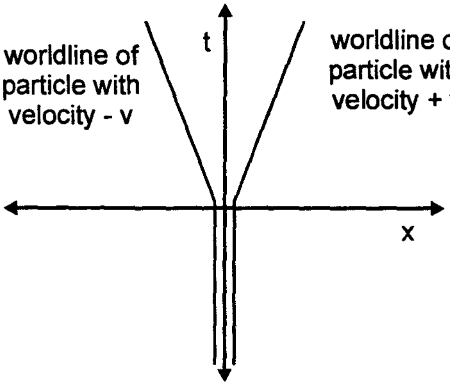

worldline of particle with velocity - v
t
worldline of particle with velocity + v
x

Fig. 7 Space Time.

To examine this question, let's consider two identical neutral particles, each of mass m_0, initially sitting side-by-side, at rest at the spatial origin of an inertial coordinate system S. Since the particles have $v = 0$, the combined mass of the two particles is presumably just $2m_0$, and their total momentum is zero. Now suppose that at the time $t = 0$ these two particle somehow push against each other, and begin moving away from the origin in opposite directions at the speed v.

The total momentum is still zero, but according to the total relativistic mass of the two particles is now

$$m_{tot} = \frac{2m_0}{\sqrt{1-v^2}}$$

At this point, from the standpoint of classical Newtonian mechanics augmented by the relativistic mass formula, we are faced with an apparent violation of the conservation of 'mass', because the relativistic mass has increased due to the acquired speed v of the particles. The apparent increase in relativistic mass of the particles is

$$2m_0\left[\frac{1}{\sqrt{1-v^2}} - 1\right] = m_0 v^2 + \frac{3}{4} m_0 v^4 +$$

Of course, based on what we've said so far, we also have a violation of the conservation of energy, because initial the two particles were are rest (zero kinetic energy), and then they acquired the speed v, and we gave no indication of the source of this energy. One possibility is that there is a massless spring compressed between the two particles, which are held together (initially) by a latch, as illustrated below.

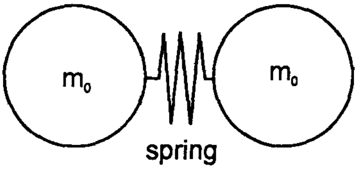


Fig. 8 Massless Spring Compressed Between the Two Particles.

When the latch is released at time $t = 0$, the spring drives the particles apart. Thus the original and final configurations have the same energy, initially stored as potential energy in the compressed spring, and later in the form of kinetic energy of the particles. According to Newtonian mechanics (augmented with relativistic mass) the final kinetic energy of the two particles is

$$K = \frac{m_0 v^2}{\sqrt{1-v^2}} = m_0 v^2 + \frac{1}{2} m_0 v^4 +$$

Thus to the second order the kinetic energy imparted to the particles equals the increase in the relativistic mass of the particles. It isn't hard to show that the energy originally contained in the compressed massless spring must contribute actual rest mass to the original configuration. In other words, the total rest mass of the two particles latched together with the compressed massless spring between them must be greater than $2m_0$.

To show this, consider the same two particles, but this time the latched pair originally has a speed u in the positive x direction, and then at time $t = 0$ the latch is released, as indicated in the spacetime diagram below.

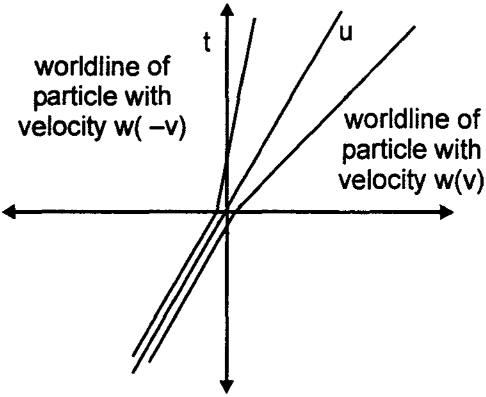

t
u
worldline of particle with velocity w(–v)
worldline of particle with velocity w(v)

Fig. 9 Velocity of Particle.

Now, a system of inertial coordinates for which the particles were originally at rest will be defined such that the two particles acquire the speed v is opposite directions when the latch is released. However, with respect to the original system of inertial coordinates, it's clear that the speeds cannot equal simply $u + v$ and $u - v$. We can say only that the final speeds (for any given u) will be $w(v)$ and $w(-v)$ for some function w, which we know must be linear fractional because these are the only entire meromorphic bijections.

Thus for any given u we have

$$w(v) = \frac{a+bv}{c+dv}$$

where a,b,c,d are functions of u. For conservation of relativistic momentum with respect to the original coordinates, we must have

$$\frac{M_0 u}{\sqrt{1-u^2}} = \frac{m_0 w(v)}{\sqrt{1-w(v)^2}} + \frac{m_0 w(-v)}{\sqrt{1-w(-v)^2}}$$

where M_0 is the total rest mass of the initial configuration of the two particles plus the massless compressed spring between them (with energy to propel the particles to the speed v when unlatched). Inserting the linear fractional form of w into this equation and simplifying, we get

$$\frac{M_0 u}{\sqrt{1-u^2}} = \frac{m_0(a+bv)}{\sqrt{(c+dv)^2-(a+bv)^2}} + \frac{m_0(a-bv)}{\sqrt{(c-dv)^2-(a-bv)^2}}$$

Without loss of generality we can set $c = 1$ (since we can divide through the linear fractional transformation by c). Also, in the case $v = 0$ we have $M_0 = 2m_0$ and the above equation reduces to

$$\frac{2m_0 u}{\sqrt{1-u^2}} = \frac{2m_0 a}{\sqrt{1-a^2}}$$

so we must have $a(u) = u$. Furthermore, we note that the denominators on the right side are equal to each other if and only if $ab = cd$, so if we set $ub = d$ the full equation becomes

$$\frac{M_0 u}{\sqrt{1-u^2}} = \frac{\dfrac{2m_0}{\sqrt{1-b^2v^2}}}{\sqrt{1-u^2}}$$

We know that $b(u)$ equals or approaches 1 as u goes to zero, so evidently the rest mass M_0 of the two particles plus the energy capable of propelling them to the speed v is not $2m_0$ but rather

$$M_0 = \frac{2m_0}{\sqrt{1-v^2}}$$

Thus conservation of momentum requires that the rest mass of the original configuration consist not just of the rest masses of the two particles, but also a contribution equal to the energy necessary to propel them apart at the speed v. We suggested an ideal massless spring to represent this energy, but the above argument applies to any form of energy (e.g., a small amount of gunpowder, accounting for the chemical energy). Hence we are forced to conclude that energy itself has inertia, and bound energy contributes to the rest mass of an object. It follows that relativistic mass is simply a measure of the energy of a system, so the conservation of energy is equivalent to the conservation of relativistic mass. Furthermore, for consistency with the relativistic mass formula and the conservation of momentum, we must have $a = d = u$ and $b = c = 1$, so the speeds (with respect to the original coordinates) of the two particles when unlatched must be given by

$$w(v) = \frac{u+v}{1+uv} \quad w(-v) = \frac{u-v}{1-uv}$$

whereas the speeds of those particles with respect to the inertial coordinate system in terms of which the original configuration was at rest are simply $+v$ and $-v$ by the definition of inertial coordinates. This expression for composition of velocities leads, in turn, to the Lorentz transformation as the relationship between relatively moving systems of inertial coordinates.

In summary, we've shown that the relativistic mass formula, combined with conservation of momentum in the Newtonian context, does indeed imply special relativity, at least to the extent of implying the equivalence of mass and energy, and the Lorentz transformation relating relatively moving systems of inertial coordinates. It might be argued that we have considered only mechanical systems, and can't conclude anything about (for example) the propagation of light from equation. However, we've seen that implies all forms of energy have inertia. For example, we could imagine a pulse of light bouncing between the mirrored surfaces of the two particles when latched together, and we would find that the energy of that pulse of light contributes to the rest mass of the bound configuration, just as did the energy of a compressed spring. Hence a pulse of light is an inertial entity, and must satisfy conservation of momentum, so it too must propagate in conformity with equation, which of course involves the constant c, defined as the speed of light. Hence we have tacitly introduced electromagnetism into the relativistic mass formula.

The usefulness (and even the validity) of the concept of 'relativistic mass', as distinct from 'rest mass', has often been debated, but it seems to be overlooked by both sides that at least some (and perhaps all) of the rest mass of typical physical entities ultimately consists of relativistic mass. For example, the 'rest mass' of a billiard ball at a given temperature consists partly of the relativistic mass of the molecules comprising the ball, since those molecules are in motion even while the ball is stationary. Also the binding energy of molecules, and various other forms of potential energy, contribute to the rest mass, so the idealized image of 'rest mass' as a primitive attribute of some localizable entity seems fundamentally misguided. The idea of invariant rest mass is perhaps best adapted to the study of sub-atomic particles, since those entities come closest to being irreducible, and hence the assignment of a primitive invariant rest mass to such entities often seems justified. However, even in this context we know an electron and a positron can annihilate each other, giving off electromagnetic energy, leaving no 'matter' at all. Of course, it remains possible to define the 'rest mass' of the total emitted electro-magnetic energy, but in so doing we are identifying a rest mass for the combination of multiple entities (photons), each of which has zero rest mass. If effect, this rest mass consists entirely of relativistic mass.

Special Relativity

An *inertial coordinate system* is a system of space and time coordinates in terms of which the resistance to acceleration of any given object at rest is the same at every location and in all directions. In other words, we define inertial coordinate systems in such a way that the inertia of material objects is homogeneous and isotropic. Homogeneity implies that every material body free of external influence moves at constant speed in a straight line, and isotropy implies that if two identical material objects initially adjacent and at rest act to repel each other, they acquire equal speeds in opposite directions. Given one inertial coordinate system we can construct infinitely many others by means of arbitrary fixed translations and spatial rotations, which leave the speed of every object unchanged. Such an equivalence class of inertial coordinate systems is called an *inertial reference frame*. It's important to recognize that the definition of an inertial reference frame not only identifies inertial motion with straight paths of constant speed, it also establishes an operational definition of simultaneity (i.e., the synchronization of times at spatially separate events), because inertial isotropy implies that we can use identical physical objects acting against each other to synchronize clocks equidistant from their centre of mass (relying on the equilibrium configurations of solid objects at rest to define distances).

Given this definition of inertial reference frames, the *principle of relativity* asserts that for any material particle in any state of motion there exists an inertial reference frame—called the *rest frame* of the particle—with respect to which the particle is instantaneously at rest (i.e., the change of the spatial coordinates with respect to the time coordinate is zero). This principle is usually extended to include *reciprocity*, meaning that for any two systems S_1 and S_2 of inertial coordinates, if the spatial origin of S_1 has velocity v with respect to S_2, then the spatial origin of S_2 has velocity $-v$ with respect to S_1. The existence of this class of reference frames, and the viability of the principles of relativity and reciprocity, are inferred from experience. Once these principles have been established, the relationship between relatively moving inertial coordinate systems can then be considered.

Let $[t,x,y,z]$ signify a system of inertial coordinates in the rest frame of particle p, and likewise let $[t',x',y',z']$ signify a system of inertial coordinates in the rest frame of a particle p' moving with speed v relative to $[t,x,y,z]$. By means of a fixed translation we can make the origins of these two coordinate systems coincide, and by a fixed spatial rotation we can spatially align the x and x' axes. For simplicity, we will consider particles and motions confined to the x, x' axes. The question naturally arises as to how these two coordinate systems are related to each other for a given relative velocity v. Since, by definition, inertial motions are straight lines with respect to both systems, the relations between two inertial coordinate systems must be linear functions of the form

$$x' = Ax + Bt,\ t' = Cx + Dt$$

for constants A,B,C,D (for a fixed v). A stationary object in the rest frame of p' has a constant value of x', so the differential $dx' = Adx + Bdt = 0$ implies $dx/dt = -B/A = v$ and hence $B = -vA$. The inverse of the above transformation is

$$x = \frac{Dx'-Bt'}{AD-BC} \quad t = \frac{-Cx'-At'}{AD-BC}$$

Evaluating the velocity of a stationary object in the frame of p with respect to the frame of p' leads to $dx'/dt' = B/D = -v$ (by reciprocity) and hence $D = A$.

Letting μ^2 denote the quantity $AD - BC$, and substituting for B and D, we have $A^2 + vAC = \mu^2$, so we have $C = (\mu^2 - A^2)/(vA)$. If we define $\alpha = A/m$, then the original transformation can be written in the form

$$x' = m\alpha\,(x - vt) \quad t' = m\alpha\left(t - \left[\frac{\alpha^2 - 1}{v^2\alpha^2}\right]vx\right)$$

and the inverse transformation has the form

$$x = \frac{\alpha}{\mu}(x' + vt') \quad t = \frac{\alpha}{\mu}\left(t' + \left[\frac{\alpha^2 - 1}{v^2\alpha^2}\right]vx'\right)$$

If we replace v with -v these two transformations are exchanged, except for the factor μ, so if we are to have spatial isotropy we must have μ equal to 1. This leaves undetermined only the expression in square brackets. (Remember that the parameter α is a function of v, but it is a constant for any fixed v.) Letting k denote this quantity, we have $[(\alpha^2-1)/(v\alpha)^2] = k$, from which we get $\alpha = 1/\sqrt{1-kv^2}$, and the general transformation can be written in the form

$$x' = \frac{x - vt}{\sqrt{1-kv^2}} \quad t' = \frac{t - kvx}{\sqrt{1-kv^2}}$$

Any two inertial coordinate systems must be related by a transformation of this form, where v is the mutual speed between them. Also, note that

$$\frac{x'}{t'} = \frac{x - vt}{t - kvx}$$

Given three systems of inertial coordinates with the mutual speed v between the first two and u between the second two, the transformation from the first to the third is the composition of transformations with parameters k_v and k_u. Letting x'', t'' denote the third system of coordinates, we have by direct substitution

$$\frac{x''}{t''} = \frac{x - \left(\dfrac{u+v}{1+k_v uv}\right)t}{\left(\dfrac{1+k_u uv}{1+k_v uv}\right)t - \left(\dfrac{k_u u + k_v v}{1+k_v uv}\right)x}$$

The coefficient of t in the denominator of the right side must be unity, so we have $k_u = k_v$, and therefore k is a constant for all v, with units of an inverse squared speed. Also, the coefficient of t in the numerator must be the mutual speed between the first and third coordinates systems. These identifications are necessary and sufficient to make the transformation be of the required form.

Now, if k is non-zero, we can make its magnitude equal to unity by a suitable choice of units for distance and time, so the only three essentially distinct cases to consider are $k = -1$, 0, or $+1$. If $k = -1$ this transformation is simply a Euclidean rotation in the xt plane through an angle $\theta =$ invtan(v). In other words, with $k = -1$ the above equations can be written in the form

$$x' = -\sin(q)\ \mathrm{t} + \cos(q)\ \mathrm{x} \qquad \mathrm{t}' = \cos(q)\ \mathrm{t} + \sin(q)\ x$$

On the other hand, with $k = 0$ the general transformation reduces to the Galilean space-time transformation, i.e., we have

$$x' = x - vt \qquad \mathrm{t}' = t$$

The remaining case is with $k = +1$, which gives the so-called Lorentzian transformation

$$\mathrm{x}' = \frac{x - vt}{\sqrt{1-v^2}} \quad \mathrm{t}' = \frac{t - vx}{\sqrt{1-v^2}}$$

We wish to determine which of these represents the correct relation between relatively moving inertial coordinate systems. The Euclidean transformation (i.e., the case $k = -1$) is easy to rule out empirically, because we cannot turn around in time as we can in space. However, it isn't as easy to distinguish empirically between the Galilean and Lorentzian transformations, especially if the value of k in ordinary units of space and time is extremely close to zero. As a result, Newtonian mechanics was based for many years on the assumption that $k = 0$.

It was not until the late 19th century that sufficiently precise experimental techniques became available to determine that the true value of k is not zero. In ordinary units it has the value $(1.11)10^{-17}$ second^2/metre^2 (which happens to equal $1/c^2$ where c is the speed at which electromagnetic waves propagate in vacuum). This implies that relatively moving systems of inertial coordinates are related according to the Lorentzian transformation. It follows that the constant-t surfaces of two relatively moving systems of inertial coordinates are skewed, although the skew is so slight it's nearly impossible to detect in most ordinary circumstances.

It was also found that Maxwell's equations of electromagnetism — which were developed without knowledge of the Lorentz transformation—are actually invariant with respect to the Lorentz transformation. Hence all the phenomena of electromagnetism and of mechanical inertia are relativistic with respect to precisely the same class of transformations. Furthermore, it was found that whatever forces are responsible for the stability of matter (which was known to be inexplicable in terms of electromagnetism alone) are also invariant with respect to the Lorentz transformation. Naturally this led to the hypothesis that all physical phenomena, including gravity and any other physical forces that may exist, are invariant with respect to the Lorentz tranformation. Subsequently this hypothesis has been confirmed for gravity, for the strong and weak nuclear forces, and for all quantum mechanical processes. In fact, no violation of Lorentz covariance in any physical phenomenon has ever been detected (despite strenuous efforts).

Notice that for any incremental interval whose components are (dt, dx) with respect to one particular system of inertial coordinates, and (dt', dx') with respect to any other system of inertial coordinates, we have

$$(dt')^2 - (dx')^2 = (dt)^2 - (dx)^2$$

This signifies that the quantity $(dt)^2 - (dx)^2$ is invariant with respect to all inertial coordinate systems. Since there exists a system of inertial coordinates with respect to which the spatial component dx is zero, it follows that the above invariant quantity is the square of the time differential dt along a path with respect to the rest frame of the path. This particular time differential is an invariant quantity, called the *proper time* of the interval (usually denoted by $d\tau$ to distinguish it from an arbitrary inertial time coordinate). It's easy to see that the inertial path between any two events has the maximum lapse of proper time. Any non-inertial path between those same two events will have a lesser lapse of proper time. In general, the lapse of proper time along the path of any physical entity corresponds precisely to the advance in the phase of the entity's quantum wave function. According to quantum mechanics, the wave function encodes everything knowable about a physical system, and there are no underlying structures or 'hidden variables', so the proper time along any interval actually *is* physical time.

When people first hear about special relativity they often wonder if it's necessary to think in terms of coordinate systems whose constant-t surfaces are skewed. They point out that it's possible to construct a set of relatively moving coordinate systems that all share a common time coordinate. However, if two relatively moving systems of coordinates share a common time parameter, they cannot both be *inertial* coordinate systems. The definition of an inertial coordinate system already imposes a specific set of constant-t surfaces for any given time axis in order to make inertia isotropic (i.e., the same in all spatial directions). We are certainly free to think in terms of non-inertial coordinate systems, but then we must be careful to remember that inertia is not isotropic with respect to such coordinate systems.

One major shortcoming with the way in which special relativity is usually taught is that inertial coordinate systems (and frames) are usually not fully defined. They are typically characterized simply as coordinate systems that are 'not accelerated', which is to say, coordinate systems in terms of which inertia is homogeneous. It's vitally important to realize that being unaccelerated is a necessary but not a sufficient condition for a coordinate system to be inertial, because there exist coordinate systems with spatial axes oblique to the time axes (relative to the inertial orientation). In such systems, all inertial motion has uniform speed in a straight line, but nevertheless the coordinates are not (in general) inertial, because although Newton's first law of motion is satisfied, his second and third laws are not (even quasi-statically). In other words, inertia is not generally isotropic with respect to oblique coordinates. For any given time axis there is a unique orientation of the spatial axes compatible with inertial isotropy. This shows the significance of Einstein's statement at the beginning of Part I of his 1905 paper on electrodynamics: "Let us take a system of coordinates in which the equations of Newtonian mechanics hold good." It is significant that he does not limit this to just Newton's first law. Of course, one result of Einstein's paper is that Newton's laws as traditionally formulated are valid only quasi-statically, but the point is that he's clearly referring to systems of coordinates in terms of which inertia is not only homogeneous but also isotropic, because without isotropy Newton's third law is not even quasi-statically valid.

Students are often told that Einstein and/or Poincare were the first to introduce operational definitions of simultaneity, but in fact there has always been an operational definition of simultaneity for inertial coordinates, because there is a unique simultaneity compatible with inertial isotropy for any given time axis. Galileo himself explained this in his "Dialogues on the Two Chief World Systems". What the discoveries of Bradley, Fizeau, Maxwell, Michelson, etc., made clear was that the propagation of electromagnetic disturbances is isotropic with respect to the same class of coordinate systems (the inertial coordinate systems) in terms of which mechanical inertia is isotropic.

In addition, they found that the value of k is not exactly zero (coincidentally at about the same time that Planck discovered that the value of h is not exactly zero), and consequently there is an invariant speed, equal to, for the set of inertial coordinate systems. From this we immediately have the concept of proper time, which is identified with the phase of the quantum wave function along any given timelike path.

Another common misconception is that we cannot assert the empirical isotropy of the one-way speed of light with respect to inertial coordinates. It's actually quite possible to demonstrate this one-way isotropy. Simply observe that two identical particles acting on each other reach any given distance from their common centre of mass coincident with two pulses of light emanating from that centre. Obviously this does not demonstrate the one-way speed of either the particles or the light pulses, but it does demonstrate that if we define a system of coordinates such that mechanical inertia is isotropic (i.e., an inertial coordinate system), then the one-way speed of light is also isotropic with respect to that system of coordinates.

Quantum Entanglement and Bell's Theorem

A simple description of the essential non-classical nature of quantum entanglement is given in Entangled Choices. For a more detailed and technical analysis, consider the quantum spin of an electron (or any spin-1/2 particle), which is always found to have either the value $\hbar+/2$ or $-\hbar/2$ in the direction of measurement, regardless of the direction we choose to measure. (This was first shown for silver atoms in the famous Stern-Gerlach experiment, and has subsequently been verified for many other kinds of particles.) Thus an electron manifests one of only two possible spin states, which we may call 'spin up' and 'spin down'. It's convenient to represent these states as orthogonal unit vectors

$$\langle \text{up} \rangle = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad \langle \text{down} \rangle \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

In general an electron's spin state, Ψ, at any instant can be represented by a linear combination of those two possible observable states. The choice of a measurement direction is equivalent to choosing a 'basis' for expressing the spin components of the state Ψ. For any particular basis we can express the state in the form

$$\Psi = c_1 \begin{bmatrix} 1 \\ 0 \end{bmatrix} + c_1 \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix}$$

where c_1 and c_2 are complex constants. These constants encode the probability that the chosen measurement will yield either result. The probability is simply the norm of the respective coefficient. For example, if $c_1 = a + ib$ and we let $c_1{}^*$ denote the complex conjugate, $a - ib$, then the probability of a measurement on this basis yielding 'spin up' is equal to the norm, i.e., the product $c_1{}^*c_1 = a^2 + b^2$. Likewise the probability of finding 'spin down' on this basis is $c_2{}^*c_2$. Since these are the only two possilities, we have

$$c_1{}^*c_1 + c_2{}^*c_2 = 1$$

As always in quantum mechanics, each possible measurement basis is associated with an operator whose eigenvalues are the possible results of the respective measurement. For a given xyz basis of orthogonal space coordinates we can represent the three principle measurements (i.e., measurements along the three axes) by the matrices

$$\hat{S}_x = \frac{\hbar}{2}\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \quad \hat{S}_y = \frac{\hbar}{2}\begin{bmatrix} 0 & -\mathrm{i} \\ \mathrm{i} & 0 \end{bmatrix} \quad \hat{S}_z = \frac{\hbar}{2}\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$$

The eigenvalues of the measurement operator corresponding to whichever measurement direction we choose determine the coefficients c_1 and c_2, which represent the probabilities of the possible outcomes.

To see how this works, suppose Ψ_1 is the initial state vector of the electron, and we decide to perform a spin measurement corresponding to a particular operator $\hat{S}$. The result is given by applying $\hat{S}$ to Ψ_1, using ordinary matrix multiplication, to give the new state vector Ψ_2 : $S\Psi_1 = \Psi_2$

Since Ψ_2 is either pure 'spin up' or pure 'spin down' in the direction of measurement represented by $\hat{S}$, it follows that a subsequent measurement in the same direction must yield the same result, so Ψ_2 must be such that: $\hat{S}\,\Psi_2 = \Psi_2$

for some constant λ (since state vectors are equivalent up to length). Thus the constant λ is an eigenvalue of the measurement operator $\hat{S}$, and Ψ_2 is the corresponding eigenvector.

The outcome of is unambiguous, because the eigenvector on the right side is the same as the eigenvector on the left side. However, since the arbitrary initial state Ψ_1 in is not in general an eigenvector of $\hat{S}$, it can yield either of the eigenvectors of $\hat{S}$. This reveals the probabilistic aspect of quantum mechanics.

The eigenvectors of $\hat{S}$ constitute a basis for the space of possible state vectors, so Ψ_1 can be expressed as a linear combination of those eigenvectors. If we let Ψ_2 and $\Psi_{2'}$ denote the eigenvectors, then we can express Ψ_1 as

$$\Psi_1 = C_1\,\Psi_2 + C_2\,\Psi_{2'}$$

Again, the norm of each complex coefficient gives the probability that $\hat{S}$ applied to Ψ_1 will lead to the respective eigenstate.

For the three principle direction spin operators $\hat{S}_x$, $\hat{S}_x$, and $\hat{S}_x$ presented above, the eigenvectors corresponding to$+\hbar/2$ or $-\hbar/2$:

$$\begin{array}{ccc} & -\hbar/2 & +\hbar/2 \\ \hat{\mathrm{S}}_{\mathrm{x}} & \frac{1}{\sqrt{2}}\begin{bmatrix}-1\\1\end{bmatrix} & \frac{1}{\sqrt{2}}\begin{bmatrix}1\\1\end{bmatrix} \\ \hat{\mathrm{S}}_{\mathrm{y}} & \frac{1}{\sqrt{2}}\begin{bmatrix}1\\-\mathrm{i}\end{bmatrix} & \frac{1}{\sqrt{2}}\begin{bmatrix}1\\\mathrm{i}\end{bmatrix} \\ \hat{\mathrm{S}} & \begin{bmatrix}1\\0\end{bmatrix} & \begin{bmatrix}1\\0\end{bmatrix} \end{array}$$

Each pair of eigenvectors constitutes a basis for the state space, so we can express the electron's state vector as a linear combination of the basis vectors for the desired measurement, and the coefficients give the probability of that measurement yielding either 'spin up' or 'spin down'. We can think of these probabilities as the projections of the initial state vector onto the orthogonal axes of the chosen measurement basis.

Of course, we aren't restricted to measurements along one of the principle axes. We can measure the spin of the electron along any spatial axis, and each such measurement is represented by an operator. We also note that these directions are purely relative to the state of the particle in question. To illustrate, suppose a stream of electrons is moving along the y axis and we perform spin measurements on these particles in the z direction, which establishes the z component of the particle's spin vector. If we filter out all those particles with 'spin down' in the z direction, we are left with a stream of particles all having 'spin up' in the z direction. Now suppose we perform a spin measurement on the remaining particles along the direction in the xz plane at an angle q with the positive z axis. In accord with the interpretation of probabilities as the projections of the state vector onto the basis axes, we infer that the measurement operator for spin in this direction is given by the projections of the x and z operators

$$\hat{S}_q = sin(\theta)\,\hat{S}_x + cos(\theta)\hat{S}_y = \frac{\hbar}{2}\begin{bmatrix} cos(\theta) & sin(\theta) \\ sin(\theta) & -cos(\theta) \end{bmatrix}$$

Naturally the eigenvalues of this operator are $+h/2$ and $-h/2$, and it's easy to show that the corresponding eigenvectors are

$$\begin{bmatrix} \cos(\theta/2) \\ sin(\theta/2) \end{bmatrix} \text{ and } \begin{bmatrix} -sin(\theta/2) \\ cos(\theta/2) \end{bmatrix}$$

respectively. Each particle subjected to this measurement will have the initial state vector indicating 'spin up' in the z direction, and we can express that initial vector as a linear combination of these basis vector.

$$\begin{bmatrix} 1 \\ 0 \end{bmatrix} = c_1 \begin{bmatrix} cos(\theta/2) \\ sin(\theta/2) \end{bmatrix} + \begin{bmatrix} -sin(\theta/2) \\ cos(\theta/2) \end{bmatrix}$$

Which implies that the coefficients are $c_1 = cos(q/2)$ and $c_2 = -sin(q/2)$. Consequently, the probabilities of 'spin up' and 'spin down' for a measurement of such a particle along the q direction are $cos(q/2)^2$ and $sin(q/2)^2$ respectively.

These quantum mechanical predictions (which have been well supported by experiment) have some remarkable implications. In the preceding example we established the initial state vector by performing a measurement in the z direction and considering only particles that yielded 'spin up' in that direction. Then we subsequently performed a measurement at an angle q relative to the positive z direction, and the probabilities of the outcomes were found to be a function of the angle q between those two measurements. An essentially equivalent experiment can be performed by examining the spins of two spin-1/2 particles emitted in opposite directions from the decay of a singlet state with zero total spin. In such a case conservation of angular momentum requires that the spin state vectors of the individual particles are precisely opposite, so if we measure the spin of one of these particles along a certain direction and find 'spin up' in that direction, then the other particle must have pure 'spin down' in that direction. Thus, by measuring the spin of one particle and reducing its state vector to one of the eigenvectors of the chosen measurement basis, we automatically 'collapse the wavefunction' of the other particle onto this same basis.

At first this process may not seem very mysterious, since it's easy to imagine that the two coupled particles are 'programmed' with opposite spins, such that they will always give opposite results when measured along any given direction. However, a careful analysis of the quantum mechanical predictions for all possible combinations of measurement angles reveals the need for a profound shift in the classical view of the world.

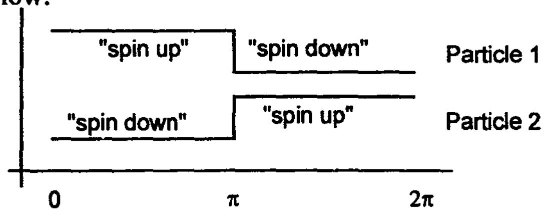
If the measurement of one particle along a fixed direction (in the *xz* plane) yields 'spin down', then the other particle is purely 'spin up' in that direction. Consequently if we perform a measurement on the other particle along a direction at an angle of θ from the first measurement, we've already seen that the probability of 'spin up' is $cos(\theta/2)^2$ and the probability of spin down in $sin(\theta/2)^2$. In a similar way we can show that if the measurement of the first particle yields 'spin up', then the other particle is purely 'spin down' along that direction, and a measurement of that other particle along a direction at an angle q relative to the first will yield 'spin up' with probability $sin(\theta/2)^2$ and spin down with probability $cos(\theta/2)^2$.

Hence the probability that the measurements of these two particles at angles differing by θ will both give the same result (both up or both down) is $sin(\theta/2)^2$, and the probability that they will yield opposite results (one up and one down) is $cos(\theta/2)^2$. The angle q between the two measurements can be expressed as $\alpha-\beta$ where α is the angle of the measurement performed on one of the particles and β is the angle of the measurement performed on the other. The two particles emitted from a singlet state are said to be entangled, because regardless of how far apart they travel before the spin measurements are made, the joint results will exhibit these *joint* probabilities.

From a classical standpoint we would imagine that each particle emerges from the singlet state with, in effect, a set of pre-programmed instructions for what spin to exhibit at each possible angle of measurement, or at least what the probability of each result should be. The usual approach to analyzing these phenomena classically is to stipulate that a particle's pre-programmed instructions for responding to a measurement at a given angle must be definite and unambiguous (rather than probabilistic) because we classically regard the two measurement angles as independent, which implies that the measurement on the 'other' particle could be at the same angle as our measurement of *this* particle, and the particles *must* give opposite results in that case. Likewise the measurement on the 'other' particle could be 180 degrees away from our measurement of this particle, and the particles *must* give equal results in that case. Of course, the individual measurements can each be either 'spin up' or 'spin down' in both of these cases, so in principle they could still be probabilistic tendancies, but classically we have no way of ensuring perfect correlation (or perfect anti-correlation) of the joint results of spacelike separated events other than by definitely pre-programming the spins of each particle for each possible measurement angle.

From this assumption it follows that the instructions to one particle are just an inverted copy of the instructions to the coupled particle. In other words, for each measurement angle from 0 to 2π the pre-programmed response to a spin measurement for one particle is the opposite of the pre-programmed response of the other particle at that angle. Furthermore, since we have perfect correlation if our measurements are at angles that differ by 180 degrees, it follows that the pre-programmed instructions for each particle are individually anti-symmetric at a phase angle of 180 degrees. For example, if a particle's programmed response for a measurement at angle α is 'spin up', then the programmed response of that same particle for a measurement at angle $\alpha + \pi$ must be 'spin down'. Hence we can fully specify the instructions to both particles by simply specifying the instructions to one of the particles for measurement angles ranging from 0 to π.

The simplest function that satisfies the conditions stated so far is constant 'spin up' over the range from 0 to π for one of the particles. This gives the total 'response programmes' of both particles shown below:



Naturally we can take the absolute orientation of this profile as arbitrary relative to our measurements. Unfortunately this simple profile doesn't give the correct joint correlations for measurement angles that differ by amounts other than 0 or 180 degrees. If we define the *correlation* exhibited by the measurements results of a set of particle pairs as the number of agreements minus the number of disagreements, all divided by the total number of pairs, then it's easy to see that the correlation yielded by pairs with the above instruction profiles (oriented randomly) varies linearly as a function of the difference between the measurement angles. The correlation is -1 if the angles differ absolutely by 0, the correlation is zero if the angles differ by $\pi/2$ (90 degrees), and the correlation is $+1$ if the angles differ by p (180 degrees). In general, for measurement angles differing by θ, the correlation is $C(\theta)=(2/\pi)|\theta|-1$. In contrast, since the quantum mechanical predictions for agreement and disagreement are $sin(\theta/2)^2$ and $cos(\theta/2)^2$ respectively, the quantum mechanical prediction for the correlation is

$$\frac{sin(\theta/2)^2-cos(\theta/2)^2}{sin(\theta/2)^2+cos(\theta/2)^2} = \frac{1-cos(\theta)}{2}-\frac{1-cos(\theta)}{2} = -cos(\theta)$$

A plot of the simple linear correlation profile and the quantum mechanical profile is shown.

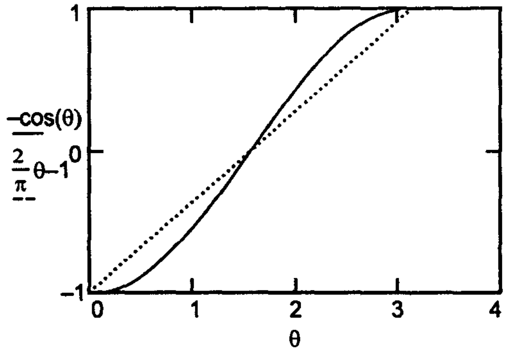


Fig. 10

These profiles agree only for measurements differing by 0, $\pi/2$, and π. For all other cases the simple pre-programmed linear model fails to match the predictions of quantum mechanics—which have been amply verified by experiments. This raises the interesting question of whether *any* pre-programmed response profile can reproduce the predictions of quantum mechanics. Suppose each particle is programmed with a more complicated profile of responses as a function of the measurement angle. In general we could partition the angular range from 0 to π into n arbitrary segments, alternating between 'spin up' and 'spin down'. Together with the perfect correlation and anti-correlation requirements at angluar differences of 0 and π, this suffices to determine the putative response profiles of both particles over the entire range from 0 to 2π. Letting $S(\alpha)$ denote the spin result (+1 for up and −1 for down) at the measurement angle a for one of the particles, and letting $s(\alpha)$ denote the spin result of the other particle, we have $S(\alpha) = -s(\alpha)$ and $S(\alpha) = -S(\alpha+\pi)$. The two results agree for measurement angles α and β if $S(\alpha)s(\beta) = +1$, and they disagree if $S(\alpha)s(\beta) = -1$. The absolute values of α and β are taken to be arbitrary, so the probability of agreement for all cases where $\beta - \alpha = \theta$ for any fixed θ is given by integrating the quantity

$$\frac{1+S(\alpha)s(\beta)}{2} = \frac{1-S(\alpha)S(\beta)}{2} = \frac{1-S(\alpha)S(\alpha+\theta)}{2}$$

from $\alpha = 0$ to π. (The result is symmetrical for $\alpha - \beta = \theta$, and for the range from π to 2π.) Therefore, equating this with the quantum mechanical result, we must have

$$\frac{1}{\pi}\int_{\alpha-0}^{\pi}\frac{1-S(\alpha)S(\alpha+\theta)}{2}d\alpha = \sin\left(\frac{\theta}{2}\right)^2 = \frac{1-\cos(\theta)}{2}$$

which implies

$$-\frac{1}{\pi}\int_{\alpha=0}^{\pi} S(\alpha)S(\alpha+\theta)d\alpha = -\cos(\theta)$$

This is just the correlation between the two measurements. If $S(\alpha) = +1$ on the entire interval from $\alpha = 0$ to π (and therefore $S(\alpha) = -1$ for all a from π to 2π), the left hand side can be split into two integrals, one extending from $\alpha = 0$ to $\pi - \theta$ with an argument of +1, and the other from $\alpha = \pi - \theta$ to π with an argument of −1. Hence the left side reduces to $(2/\pi)\theta - 1$, as we saw earlier, and this does not match the quantum mechanical prediction. In particular, we note that the rate of change of the quantum correlation near $\theta = 0$ is zero, whereas the rate of change of the correlation for based on our simple S function has a positive slope at $\theta = 0$. This is because the increase in correlation is proportional to the increase in θ, arising from the transition at $\alpha = \pi - \theta$. Clearly this rules out *every* fixed S function, because if any more transitions are added to the pre-programmed instructions, the slope of the resulting correlation will increase proportionately.

The solid bars at the bottom signify the angular regions where ‘spin up’ has been pre-programmed for this particle. Since there are three times as many transitions, the slope of correlation versus angle is three times as great as in the basic case considered previously. This is true regardless of how these transitions are distributed.

This S function has nine transitions, so the initial slope is nine times as great as the baseline case. We're forced to conclude that no pre-programmed set of results for the range of possible measurement angles can possibly reproduce the quantum mechanical predictions (and experimental results). Equation does have the formal solution

$$S(\alpha) = \sqrt{2}\, sin\,(\alpha)$$

but of course this S function isn't restricted to the discrete values +1 and −1, so it can't be directly interpreted as a spin indicator.

Many treatments of quantum entanglement discuss 'hidden variables', and they derive inequalities involving the correlations as functions of these putative variables. In a sense, these variables are redundant, because the essence of the so-called 'local realistic' premise is that each particle emerges from the singlet state with definte instructions for the spin it will exhibit for each possible measurement angle. This implies that the only relevant free variable is the reference orientation, since the instruction profile can be oriented in any direction (relative to the measurements) with uniformly distributed probability. Integrating the product of spins over the entire angular range with a fixed difference angle must yield the correlation of the two spin measurements.

It might be argued that we need not assume any single particle exhibits the quantum mechanical probabilities, because these probabilities can only be evaluated by performing multiple measurements, so we might imagine that it's necessary to evaluate the results over a sequence of particle pairs, rather than just a single pair. Moreover, we might think that this sequence could possess 'memory', making the results of successive pairs dependent. However, in principle we could prepare a large number of particle pairs in an identical way, in spacelike-separated locations, and perform measurements on the pairs independently (again based on the naive conventional concept of local realism). According to quantum mechanics we would still expect the combined results to satisfy the same correlations. This implies that each particle pair must embody the overall propensities that we expect to find manifested in multiple trials.

Nevertheless, it's worthwhile to review the traditional derivation of Bell's inequality based on the premise that each particle emerging from the singlet state possesses a definite spin as a function of the measurement angle and some 'hidden variable' m associated with that pair. We assume an arbitrary probability density function $\delta(\mu)$, which gives the probability density for any given pair of particles being produced with that value of μ. The observed spins of the two particles are denoted by $S(\alpha, \mu)$ and $s(\beta, \mu)$, where α and β are the angles of the respective measurements. As before, the perfect anti-correlation of the two particles measured at the same angle (according to quantum mechanics) implies that $S(\alpha, \mu) = -s(\alpha, \mu)$. (In our previous discussion the implicit 'hidden variable' was simply the absolute reference angle for the measurements, with a uniform probability density over the range from $\mu = 0$ to 2π, but we can argue more generally for any hidden variable with any probability density over any range.)

Now we consider the pairwise spin correlations for three different angles, α, β, and ϕ. If one member of each coupled pair of particles has its spin measured at the angle a and the other is measured at the angle ϕ, the correlation between the measurements is given by

$$C\,(\alpha,\,\phi) = \int S(\alpha,\mu)s(\phi,\mu)\delta(\mu)\,d\mu$$

where the integral is evaluated over the full range of m. Likewise if one particle is measured at b and the other at *f*, the correlation is

$$C\,(\beta,\,\phi) = \int S(\beta,\mu)s(\phi,\mu)\delta(\mu)\,d\mu$$

Taking the difference between these two, we have

$$C(\alpha,\phi)-C(\beta,\phi)=\int s(\phi,\mu)\left[S(\alpha,\mu)-S(\beta,\mu)\right]\delta(\mu)\,d\mu$$

Since every spin value is either +1 or −1, it follows that $S(\alpha, \mu)^2 = 1$, so the value of $S(\beta, \mu)$ in this expression can be written as $S(\alpha, \mu)^2$ $S(\beta, \mu)$, and we can then factor out from the square brackets to give

$$C(\alpha,\phi)-C(\beta,\phi)=\int s(\phi,\mu)S(\alpha,\mu)\left[1-S(\alpha,\mu)S(\beta,\mu)\right]\delta(\mu)\,d\mu$$

Now we take the absolute value of both sides, and make use of the obvious fact that the absolute value of a definite integral of a function is less than or equal to the integral of the absolute value of the function, we have the inequality

$$\left|C(\alpha,\phi)-C(\beta,\phi)\right| \le \int\left|s(\phi,\mu)S(\alpha,\mu)\left[1-S(\alpha,\mu)S(\beta,\mu)\right]\delta(\mu)\right|d\mu$$

Conventionally, a probability density function always has a non-negative real value, so we stipulate that d(m) is non-negative. Also, since the product $S(\alpha, \mu)\ S(\beta, \mu)$ is always +1 or −1, the quantity in the square brackets is always either +2 or 0. The leading factors, $s(\phi, \mu)\ S(\alpha, \mu)$, always equal +1 or −1, so they determine the sign of the integrand, which is discarded by taking the absolute value. Hence the expression reduces to

$$\left|C(\alpha,\phi)-C(\beta,\phi)\right| \le \int\left[1-S(\alpha,\mu)S(\beta,\mu)\right]\delta(\mu)d\mu$$

Replacing S(β,μ) by −s(β,μ), and splitting the integral into two parts, give

$$\left|C(\alpha,\phi)-C(\beta,\phi)\right| \le \int\delta(\mu)d\mu + \int S(\alpha,\mu)s(\beta,\mu)\delta(\mu)d\mu$$

The first integral is simply 1, since $\delta(\mu)$ is a probability density function, and the second integral is the correlation for two entangled particles measured at the angles α and β. Hence we can subtract this from both sides to give Bell's inequality

$$\left|C(\alpha,\phi)-C(\beta,\phi)\right| - C(\alpha,\beta) \le 1$$

We can evaluate the left hand function assuming the quantum mechanical correlation $C(x,y) = -cos(x-y)$. This gives $|cos(\alpha-\phi) - cos(\beta-\phi)| + cos(\alpha-\beta)$. The coloured regions represent that cases where this quantity exceeds 1, and therefore violates the above inequality. The maximum violations occur in the four white regions, where the function attains the value of about 1.5.

A different approach to analyzing quantum entanglement is to assume that each particle is pre-programmed not with a definite result for each measurement angle, but with a probability of yielding specific results. Needless to say, this model will be incompatible with the premise of independence of measurement angles, because a probabilistic model can't automatically enforce the perfect correlations and anti-correlations predicted by quantum mechanics at $\theta = 0$ and π. Still, the previous discussion has shown that the naive premises of 'local realism and free choice' are untenable anyway, so it's worthwhile to examine a probabilistic scheme.

Thus we imagine that the first particle contains a function $f(\alpha)$ that represents the probability of yielding spin up for the measurement angle α. Likewise the second particle is programmed with the function $g(\beta)$ that represents the probability of yielding ‘spin up’ for the measurement angle β. In these terms we can express the probabilities of agreement and disagreement as

$$\begin{aligned}\Pr\{\text{agree}\} &= f(\alpha)g(\beta)+\left[1-f(\alpha)\right]\left[1-g(\beta)\right] \\ &= \sin\left(\frac{\alpha-\beta}{2}\right)=\frac{1-\cos(\alpha-\beta)}{2}\end{aligned}$$

$$\begin{aligned}\Pr\{\text{disagree}\} &= f(\alpha)\left[1-g(\beta)\right]+\left[1-f(\alpha)\right]g(\beta) \\ &= \cos\left(\frac{\alpha-\beta}{2}\right)^2=\frac{1+\cos(\alpha-\beta)}{2}\end{aligned}$$

If we double the ‘disagree’ equation and subtract 1 from both sides we get

$$\left[1-2f(\alpha)\right]\left[1-2g(\beta)\right]=1-2\cos\left(\frac{\alpha-\beta}{2}\right)^2 = -\cos(\alpha-\beta)$$

Incidentally, if we expand the right-hand side using the trigonometric angle addition formula, we get the interesting expression

$$\left[1-2f(\alpha)\right]\left[1-2g(\beta)\right]+\sin(\alpha)\sin(\beta)+\cos(\alpha)\cos(\beta)=0$$

which could be regarded as an ‘inner product’ of vectors with the components of the form $sin(\alpha)$, $cos(\alpha)$, and $1-2f(a)$.

Returning to the original ‘disagree’ equation, if we assume the functions f and g individually possess derivatives, and if α and β are independent variables (as is classically the case for two freely chosen measurement angles), we can take the partial derivative with respect to β at any constant α to give

$$\sin(\alpha-\beta) = 2g(\beta)\,[1-2f(\alpha)]$$

Solving for $1-2f(\alpha)$, substituting into equation, and re-arranging terms gives

$$\frac{2\dot{g}(\beta)}{1-2g(\beta)} = -\tan(\alpha-\beta)$$

Similarly we can derive the relation

$$\frac{2\dot{f}(\beta)}{1-2f(\beta)} = \tan(\alpha-\beta)$$

It's interesting that the expressions on the left side of these equations are formally the same as the queing theory 'transition rates' of the functions *2g* and *2f* if we identify these functions with the state probabilities and the arguments α and β with the model time.

In any case, notice that these equations theoretically imply that each of the two entangled particles contains enough information to enable someone to compute the measurement angle at the *other* particle. For example, the second equation can be solved for β to give

$$\beta = \alpha - \text{invtan}\left(\frac{2\dot{f}(\alpha)}{1-2f(\alpha)}\right)$$

Hence, assuming the particle contains the complete function *f*, and assuming this function possesses derivatives, it follows that the measurement angle β at the *other* particle is a deterministic function of the measurement angle α at *this* particle. This contradicts the premise that α and β are independent variables, so the only possible conclusion is that one or more of our basic assumptions were wrong. One questionable assumption was that not only does each particle 'contain' its respective function (*f* or *g*), but that these functions are differentiable. This need not be the case. Continuous but nowhere-differentiable functions are well-known in mathematics, and we could postulate that *f* and *g* are such functions.

On the other hand, if we rule out such exotic functions, it seems that the only remaining possibility is that, even though we may believe we are free to choose the two separate measurement angles 'freely' and independently, they are actually deterministically linked. In other words, not only are the particles entangled, but so are our 'free' choices of measurement bases for those particles. This analysis suggests a much more deterministic universe than is commonly conceived, regardless of whether it's possible to extract the information about f and its derivative from the particle. (These 'free' choices may even be spacelike separated, but of course this does not imply any violation of special relativity, because it is only necessary to assume that the choice events share a common event in their causal pasts, which is always the case provided we are willing to go back far enough.)

Suppose we differentiate equation again with respect to β at constant a to give

$$-\cos(\alpha-\beta) \;=\; 2\ddot{g}(\beta)\left[1-2f(\alpha)\right]$$

The left hand side is equal to $(1-2f)(1-2g)$, so we can make this substitution, cancel the factor of $(1-2f)$, and re-arrange terms to arrive at $\ddot{g}(\beta)+g(\beta)=\frac{1}{2}$

This has the particular solution $g(\beta)=1/2$, and we can add this to the homogeneous solution to give the total general solution

$$g(\beta)=\frac{1}{2}+C_1e^{i\beta}+C_2e^{-i\beta}$$

Since this function signifies the probability of 'spin up', we expect it to be real-valued and in the range from 0 to 1. Hence we can set $C_1 = C_2 = 1/4$ to give

$$g\,(\beta)=\frac{1+\cos(\beta)}{2}$$

Applying the same analysis to $f(\alpha)$ and setting $C_1 = C_2 = -1/4$ gives

$$f(\alpha)=\frac{1-\cos(\alpha)}{2}$$

If we then substitute these expressions back into the original 'agree' and 'disagree' formulas, we find that in both cases they reduce to

$$\cos(\alpha-\beta) \;=\; \cos(\alpha)\cos(\beta)$$

Comparing this with the trigonometric identity

$$\cos(\alpha-\beta)=\cos(\alpha)\cos(\beta)+\sin(\alpha)\sin(\beta)$$

we see that it is not generally satisfied unless α or β is a multiple of π, in which case it is trivially satisfied. One way around this impasse is to postulate slightly more general functions for f and g, namely,

$$f(\alpha)=\frac{1+\cos(\alpha+\alpha_0)}{2}\qquad g(\beta)=\frac{1+\cos(\beta+\beta_0)}{2}$$

where α_0 and β_0 are constant phase angles for a given pair of particles. Substituting these functions into the basic 'agree' and 'disagree' equations gives the single condition

$$\left[1+\cos(\alpha_0)\cos(\beta_0)\right]\cos(\alpha)\cos(\beta) - \left[\sin(\alpha_0)\cos(\beta_0)\right]\sin(\alpha)\cos(\beta)$$
$$-\left[\cos(\alpha_0)\sin(\beta_0)\right]\cos(\alpha)\sin(\beta) + \left[1+\sin(\alpha_0)\sin(\beta_0)\right]\sin(\alpha)\sin(\beta)=0$$

This can also be expressed as a null line element

$$\begin{bmatrix}\cos(\alpha) & \sin(\alpha)\end{bmatrix}\begin{bmatrix}1+\cos(\alpha_0)\cos(\beta_0) & -\sin(\alpha_0)\cos(\beta_0) \\ -\cos(\alpha_0)\sin(\beta_0) & 1+\sin(\alpha_0)\sin(\beta_0)\end{bmatrix}\begin{bmatrix}\cos(\beta) \\ \sin(\beta)\end{bmatrix}=0$$

Where the determinant of the coefficient matrix is $1 + \cos(\alpha_0 - \beta_0)$. It can also be expressed in the form of a Mobius transformation between the tangents of α and β as follows:

The similarity parameter of the transformation is $1 - \cos(\alpha_0 - \beta_0)$, whereas the normalized trace is tan $((\alpha_0 - \beta_0)/2)$.

Huygens' Principle

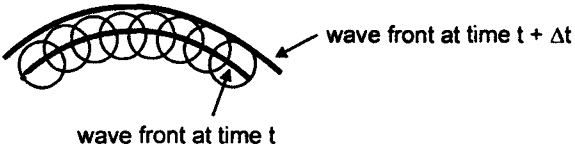


Fig. 11

This drawing depicts the propagation of the wave 'front', but Huygens' principle is understood to apply equally to any locus of constant phase (not just the leading edge of the disturbance), all propagating at the same characteristic wave speed. This implies that a wave doesn't get 'thicker' as it propagates, i.e., there is no diffusion of waves. For example, if we turn on a light bulb for one second, someone viewing the bulb from a mile away will see it 'on' for precisely one second, and no longer. Similarly, the fact that we see sharp images of distant stars and galaxies is due to Huygens' principle. However, it's worth noting that this principle is valid only in spaces with an odd number of dimensions. If we drop a pebble in a calm pond, a circular wave on the two-dimensional surface of the pond will emanate outward, and if Huygens' principle was valid in two dimensions, we would expect the surface of the pond to be perfectly quiet both outside *and inside* the expanding spherical wave. But in fact the surface of the pond inside the expanding wave (in this two-dimensional space) is *not* perfectly calm, its state continues to differ slightly from its quiescent state even after the main wave has passed through. This excited state will persist indefinitely, although the magnitude rapidly becomes extremely small. The same occurs in a space with any even number of dimensions. Of course, the leading edge of a wave always propagates at the characteristic speed *c*, regardless of whether Huygens' principle is true or not. In a sense, Huygens' principle is more significant for what it says about what happens *behind* the leading edge of the disturbance. Essentially it just says that all the phases propagate at the same speed.

From this simple principle Huygens was able to derive the laws of reflection and refraction, but the principle is deficient in that it fails to account for the directionality of the wave propagation in time, i.e., it doesn't explain why the wave front at time $t + Dt$ in the above figure is the upper rather than the lower envelope of the secondary wavelets.

Why does an expanding spherical wave continue to expand outward from its source, rather than re-converging inward back toward the source? Also, the principle originally stated by Huygens does not account for diffraction. Subsequently, Augustin Fresnel elaborated on Huygens' principle by stating that the amplitude of the wave at any given point equals the superposition of the amplitudes of all the secondary wavelets at that point. The Huygens-Fresnel principle is adequate to account for a wide range of optical phenomena, and it was later shown by Gustav Kirchoff how this principle can be deduced from Maxwell's equations. Nevertheless, it does not actually resolve the question about 'backward' propagation of waves, because Maxwell's equations themselves theoretically allow for advanced as well as retarded potentials. It's customary to simply discount the advanced waves as 'unrealistic', and to treat the retarded wave as if it was the unique solution, although there have occasionally been interesting proposals, such as the Feynman-Wheeler theory, that make use of both solutions.

Incidentally, as an undergraduate, Feynman gave a seminar on this 'new idea' at Princeton. Among the several 'monster minds' in attendance was Einstein, to whom the idea was not so new, because 30 years earlier Einstein had debated the significance of the advanced potentials with Walther Ritz. In any case, the Huygens-Fresnel principle has been very useful and influential in the field of optics, although there is a wide range of opinion as to its scientific merit. Many people regard it as a truly inspired insight, and a fore-runner of modern quantum electro-dynamics, whereas others dismiss it as nothing more than a naive guess that sometimes happens to work.

Melvin Schwartz wrote: Huygens' principle tells us to consider each point on a wavefront as a new source of radiation and add the 'radiation' from all of the new 'sources' together. Physically this makes *no* sense at all. Light does not emit light; only accelerating charges emit light. Thus we will begin by throwing out Huygens' principle completely; later we will see that it actually does give the right answer for the wrong reasons.

Whether we have now actually found the true 'reason' for the behaviour of light is debatable, and ultimately every theory is based on some fundamental principle(s), but it's interesting how widely the opinions on various principles differ. It could be argued that the 'path integral' approach to quantum field theory—according to which every trajectory through every point in space is treated equivalently as part of a possible path of the system—is an expression of Huygens' principle. It's also worth reflecting on the fact that the quantum concept of a photon necessitates Huygens' principle, so evidently quantum mechanics can work only in space with an odd number of dimensions.

Setting aside these weighty considerations, it's interesting to review the mathematical content of Huygens' original principle. The usual wave equation for a scalar field $y(x_1,x_2,..,x_n,t)$ in n space and 1 time dimension is

$$\frac{\partial^2\psi}{\partial x_1^2}+\frac{\partial^2\psi}{\partial x_2^2}+\ldots+\frac{\partial^2\psi}{\partial x_n^2}=\frac{\partial^2\psi}{\partial t^2}$$

(We've chosen units of time and space so that the wave velocity is 1.) If we consider a spherically symmetrical wave we have $\psi = \psi(r, t)$ where $r^2 = x_1^2\ x_2^2 + \ldots + x_n^2$

For future reference, notice that

$$\frac{\partial r}{\partial x_1}=\frac{x_j}{r} \text{ and } \frac{\partial^2 r}{\partial x_j^2}=\frac{r^2-x_j^2}{r^3}$$

for every index j from 1 to n. It follows that

$$\sum_{j=1}^{n}\left(\frac{\partial r}{\partial x_j}\right)^2=1 \text{ and } \sum_{j=1}^{n}\left(\frac{\partial^2 r}{\partial x_j^2}\right)=\frac{n-1}{r}$$

Returning to the basic wave equation, and assuming ψ is strictly a function of r and t, we have the following partial derivatives with respect to each of the space variables:

$$\frac{\partial \psi}{\partial x_j}=\frac{\partial \psi}{\partial r}\frac{\partial r}{\partial x_j} \quad \frac{\partial^2 \psi}{\partial x_j^2}=\frac{\partial \psi}{\partial r}\frac{\partial^2 r}{\partial x_j^2}+\frac{\partial r}{\partial x_j}\frac{\partial^2 \psi}{\partial x_j \partial r}$$

Since partial differentiation is commutative, the second factor in the last term of the right-hand equation can be written as

$$\frac{\partial^2 \psi}{\partial r \partial x_j}=\frac{\partial}{\partial r}\left(\frac{\partial \psi}{\partial r}\frac{\partial r}{\partial x_j}\right)=\frac{\partial \psi}{\partial r}\frac{\partial^2 r}{\partial r \partial x_j}+\frac{\partial r}{\partial x_j}\frac{\partial^2 \psi}{\partial r^2}$$

Now, since $\frac{\partial^2 r}{\partial r \partial x_j}=\frac{\partial}{\partial x_j}\left(\frac{\partial r}{\partial r}\right)=0$

the preceding mixed partial is simply

$$\frac{\partial^2 \psi}{\partial r \partial x_j}=\frac{\partial r}{\partial x_j}\frac{\partial^2 \psi}{\partial r^2}$$

Substituting back into the expression for the second partial derivative of ψ with respect to x_j, we have

$$\frac{\partial^2 \Psi}{\partial x_j^2}=\frac{\partial \Psi}{\partial r}\frac{\partial^2 r}{\partial x_j^2}+\left(\frac{\partial r}{\partial x_j}\right)^2\frac{\partial^2 \psi}{\partial r^2}$$

Summing all these partials for $j = 1$ to n gives

$$\sum_{j=1}^{n}\frac{\partial^2 \psi}{\partial r}=\frac{\partial \psi}{\partial r}\sum_{j=1}^{n}\frac{\partial^2 r}{\partial x_j^2}+\frac{\partial^2 \psi}{\partial r^2}\sum_{j=1}^{n}\left(\frac{\partial r}{\partial x_j}\right)^2$$

$$=\left(\frac{n-1}{r}\right)\frac{\partial \psi}{\partial r}+\frac{\partial^{2\psi}}{\partial r^2}$$

Hence the spherically symmetrical wave equation in n spatial dimensions can be written as

$$\frac{\partial^2 \psi}{\partial r^2} + \left(\frac{n-1}{r}\right)\frac{\partial \psi}{\partial r} = \frac{\partial^2 \psi}{\partial t^2}$$

Now suppose we define a new scalar field by the relation $\phi(r,t) = r^k \psi(r,t)$ where k is some fixed constant. The partial derivative of this scalar field with respect to r are

$$\frac{\partial \phi}{\partial \mathrm{r}} = r^k \frac{\partial \psi}{\partial} + kr^{k-1}\psi$$

$$\frac{\partial^2 \phi}{\partial \mathrm{r}^2} = r^k \frac{\partial^2 \psi}{\partial r^2} + 2kr^{k-1}\frac{\partial \psi}{\partial r} + k(k-1)r^{k-2}\psi$$

Notice that if we set $k = (n-1)/2$, and if we divide through this second partial by r^k, we have

$$\frac{1}{r^{(n-1)/2}}\frac{\partial^2 \phi}{\partial r^2} = \frac{\partial^2 \psi}{\partial r^2} + \left(\frac{n-1}{r}\right)\frac{\partial \psi}{\partial r} + \frac{(n-1)(n-3)}{4r^2}\psi$$

This is nearly the same as the left-hand side of the spherically symmetrical wave equation, except for the last term. Hence we can write the wave equation in the form

$$\frac{1}{r^{(n-1)/2}}\frac{\partial^2 \phi}{\partial r^2} - \frac{(n-1)(n-3)}{4r^2}\psi = \frac{\partial^2 \psi}{\partial t^2}$$

Furthermore, we can multiply through by $r^k = r^{(n-1)/2}$ to put this in the equivalent form

$$\frac{\partial^2 \phi}{\partial r^2} - \frac{(n-1)(n-3)}{4r^2}\phi = \frac{\partial^2 \phi}{\partial t^2}$$

If n equals 1, meaning that we have just a single space dimension, then $r = x_1$ and $f = y$, so we expect the second term on the left hand side to vanish identically, as indeed it does, leaving us with just the original one-dimensional wave equation, with the well-known general solution

$$\psi(r, t) = f(r - t) + g(r + t)$$

For arbitrary functions f and g. However, we might not have anticipated that the second term in the transformed wave equation also vanishes if n equals 3, i.e., in the case of *three* spatial dimensions. In this case the spherically symmetrical wave equation once again reduces to a one-dimensional wave equation, although in the modified wave function $f = r\psi$. Hence the general solution in three space dimensions is

$$\psi\ (r,\ t) = \frac{f(r-t)}{r} + \frac{g(r+t)}{r}$$

The fact that this solution is divided by r signifies that the magnitude of the wave tends to drop as r increases (unlike the one-dimensional case, in which a wave would theoretical propagate forever with undiminished strength). Focusing on just the 'retarded' component of the wave, $f(r - t)/r$, the fact that the time parameter t appears only in the difference $r - t$ implies that the (attenuated) wave propagates in time with a phase velocity of precisely 1, because for any fixed phase b we have $r - t = \beta$ and so dr/dt for this phase point is 1. Consequently if f is a single pulse, it will propagate outward in a spherical shell at precisely the speed 1, i.e., on the light cone. Conversely, it can be shown that the wave function at any point in space and time is fully determined by the values and derivatives of that function on the past light cone of the point. Any wave equation for which this is true (i.e., for which disturbances propagate at a single precise speed) is said to satisfy Huygens' principle. The connection with Huygens' original statement about secondary wavelets is that each wavelet—with the same speed as the original wave—represents a tiny light cone at that point, and Huygens' principle asserts that light is confined to those light cones.

It's worth noting that in the above derivation we were able to reduce the polar wave equation to a simple one-dimensional equation by taking advantage of the fact that an unwanted term vanished when the number of space dimensions is $n = 3$ (or $n = 1$). For the case of two dimensional space this doesn't work (nor would it work with four space dimensions). We can still solve the wave equation, but the solution is not just a simple spherical wave propagating with unit velocity. Instead, we find that there are effectively infinitely many velocities, in the sense that a single pulse disturbance at the origin will propagate outward on infinitely many 'light cones' (and sub-cones) with speeds ranging from the maximum down to zero. Hence if we lived in a universe with two spatial dimensions (instead of three), an observer at a fixed location from the origin of a single pulse would 'see' an initial flash but then the disturbance 'afterglow' would persist, becoming less and less intense, but continuing forever, as slower and slower subsidiary branches arrive. (It's interesting to compare and contrast this 'afterglow' with the cosmic microwave background radiation that we actually do observe in our 3+1 dimensional universe. Could this glow be interpreted as evidence of an additional, perhaps compactified, spatial dimension? What would be the spectrum of the glow in a non-Huygensian universe? Does curvature of the spatial manifold affect Huygens' principle?)

It turns out that Huygens' principle applies only with one time dimension and n = 3, 5, 7.., or any odd number of space dimensions, but not for any even number of space dimensions. (The case n = 1 is degenerate, because a pulse has only one path to take.) To see why, let's return to the general spherically symmetrical wave equation in n space dimensions

$$\frac{\partial^2\psi}{\partial r^2}+\left(\frac{n-1}{r}\right)\frac{\partial\psi}{\partial r}=\frac{1}{c^2}\frac{\partial^2\psi}{\partial t^2}$$

and consider a solution of the form $\psi(r, t) = f(r)g(t)$. (Naturally not all solutions are separable in this way, but since the wave equation is linear, we can construct more general solutions by summing a sufficient number of solutions of the separable form $f(r)g(t)$.) Inserting this into the wave equation and expanding the derivatives by the product rule gives

$$g\frac{d^2f}{dr^2}+\left(\frac{n-1}{r}\right)g\frac{df}{dr}=f\frac{1}{c^2}\frac{d^2g}{dt^2}$$

Dividing through by fg gives

$$\frac{1}{f}\frac{d^2f}{dr^2}+\frac{1}{f}\left(\frac{n-1}{r}\right)\frac{df}{dr}=\frac{1}{gc^2}\frac{d^2g}{dt^2}$$

Notice that the left hand side is strictly a function of r, and the right hand side is strictly a function of t.Since r and t are independent variables, the left and right sides must both equal a constant, which we will denote by κ. Hence we have two separate ordinary differential equations

$$\frac{1}{f}\frac{d^2f}{dr^2}+\frac{1}{f}\left(\frac{n-1}{r}\right)\frac{df}{dr}=\kappa \qquad\qquad \frac{1}{gc^2}\frac{d^2g}{dt^2}=\kappa$$

If κ is positive or zero the right hand equation gives 'run-away' solutions for g(t), whereas if κ is negative we can choose scaling so that $\kappa = -1$ and then $g(t)$ satisfies the simple harmonic equation, whose solutions include functions of the form $sin(ct)$ and $cos(ct)$. The left hand equation can be re-written in the form

$$r\frac{d^2f}{dr^2}+(n-1)\frac{df}{dr}+rf=0$$

If we multiplied this through by r, it would be in the form of what is called Bessel's equation, named after Friedrich Wilhelm Bessel, the German astronomer who (incidentally) was the first person to determine the distance to a star. In 1838 he determined the distance to the star called '61 Cygni' based on the parallax as viewed from the Earth at six-month intervals. Bessel functions are solutions of a standard Bessel equation, just as the ordinary trigonometric functions, sine and cosine, are solutions of the differential equation $y'' + y = 0$.

To solve the above equation we can assume a series solution of the form

$$f(r) = c_0 r^q + c_1 r^{q+1} + c_2 r^{q+2} + \ldots$$

for some integer q (which may be positive, negative, or zero) such that c_0 is non-zero. The derivatives of this function are

$$\frac{df}{dr} = qc_0 r^{q-1} + (q+1)c_1 r^q + (q+2)c_2 r^{q+1} + \ldots$$

$$\frac{d^2f}{dr^2} = q(q-1)c_0 r^{q-2} + (q+1)qc_1 r^{q-1}$$

$$+ (q+2)(q+1)c_2 r^q + \ldots$$

Substituting these into the differential equation, and collecting terms by powers of r, we get

$$c_0\left[q(q-1)+(n-1)q\right]r^{q-1} + c_1\left[(q+1)q+(n-1)(q+1)\right]r^q$$

$$+\left[\{(q+2)(q+1)+(n-1)(q+2)\}c_2 + c_0\right]r^{q+1}$$

$$+\left[\{(q+3)(q+2)+(n-1)(q+3)\}c_3 + c_1\right]r^{q+2} + \ldots$$

The coefficient of each power of r must vanish, and since c_0 is non-zero, the expression for the first coefficient implies $q(q-2+n) = 0$. This is called the indicial equation, because it determines the acceptable value(s) of q. In this case we must have either $q = 0$ or else $q = 2 - n$. If $q = 0$ then the coefficient of r^q equals $c_1(n-1)$, so either $n = 1$ or else $c_1 = 0$. On the other hand, if $q = 2 - n$, then the coefficient of r^q equals $c_1(3-n)$, so either $n = 3$ or else again $c_1 = 0$. We've already seen that the original differential equation has a particularly simple analytical solution when n (the number of space dimensions) equals either 1 or 3, so we need not consider them here. For all other value of n, we *must* have $c_1 = 0$. (Of course, even with $n = 1$ or 3, we are free to set $c_1 = 0$.)

Now, examining the coefficients of the higher powers of r, we see that in general the coefficient of r^{q+m} is of the form

$$\{(q+m+1)(q+m)+(n-1)(q+m+1)\}c_{m+1}+c_{m-1}$$

Inserting $q = 0$, setting the overall coefficient to zero, and solving for c_{m+1} gives

$$c_{m+1} = \frac{-c_{m-1}}{(m+1)(m+n-1)}$$

for $m = 0, 1, 2, \ldots$ Since c_0 is, by definition, the first non-zero coefficient, it follows that $c{-1}$ is zero, and therefore $c_1 = 0$. Moreover, applying the above formula recursively, we see that all the c_j coefficients for odd indices j must vanish. On the other hand, the coefficients with even indices are given recursively by

$$c_2 = c_0\frac{-1}{4(n+2)} \qquad c_4 = c_2\frac{-1}{4(n+2)} \qquad c_6 = c_4\frac{-1}{6(n+4)}$$

and so on. Notice that if $n = 1$ the denominators are 1×2, 3×4, 5×6, ..., etc., so the general non-zero coefficient of the solution can be written simply as

$$c_{2j} = c_0\,\frac{(-1)^j}{(2j)!}$$

giving the solution

$$f(r) = c_0\left(1-\frac{1}{2!}r^2+\frac{1}{4!}r^4-\ldots\right)$$

Hence the solution is simply $f(r) = c_0 \cos(r)$. Recall that $g(t)$ has solutions of the form $\cos(ct)$ and $\sin(ct)$, and we can create a solution given by the sum of two separable solutions, so, for example, one solution is

$$\begin{aligned}\psi(\mathrm{r,t}) &= f_1(r)g_1(ct) + f_2(r)g_2(ct)\\ &= \cos(r)\cos(ct) + \sin(r)\sin(c) = \cos(r-ct)\end{aligned}$$

Similarly if $n = 3$ the denominators of the recursive formulas are $2\cdot 3$, $4\cdot 5$, $6\cdot 7$, ..., etc., so the general non-zero coefficient is

$$c_{2j} = c_0 \frac{(-1)^j}{(2j+1)!}$$

giving the solution

$$f(r) = \frac{c_0}{r}\left(r - \frac{1}{3!}r^3 + \frac{1}{5!}r^5 - \ldots\right)$$

so in this case we have $f(r) = c_0\, sin(r)/r$. Combining this with suitable solutions for g(t) as in the case of $n = 1$, we can arrive at overall solutions such as $\psi(r, t) = cos(r-ct)/r$. This shows (again) that the cases of 1 and 3 spatial dimensions lead to especially simple solutions.

In general, for arbitrary positive integer n, the coefficient c_{2j} is of the form

$$c_{2j} = c_0 \frac{(-1)^j}{(2)(4)(6)\ldots(2j)\left[(n)(n+2)\ldots(n+2)(j-1)\right]}$$

Notice that for $n = 1$ the factors in the square brackets are consecutive odd integers, and they can be interleaved between the consecutive even integers to give a pure factorial product. Likewise for $n = 3$ the odd and even factors can be interleaved to give a pure factorial product. For higher odd integers we can interleave the factors in the same way, although there will be a fixed number of leading even factors and the same number of trailing odd factors that don't overlap. For example, with $n = 7$ the coefficient c_{12}, after re-arranging the six even and six odd factors in the denominator

$$\frac{(-1)^6 c_0}{(2)(4)\cdot\left[(6)(7)(8)(9)(10)(11)(12)(13)\right]\cdot(15)(17)}$$

Taking advantage of this interleaving, we can express the general coefficient (for sufficiently large j) with odd $n > 3$ in the form

$$c_{2j} = \frac{c_0(n-1)!}{2^{(n-3)/2}\left(\frac{n-3}{2}\right)!}\frac{1}{(2j+3)(2j+5)\ldots(2j+n-2)}$$

$$\frac{(-1)^j}{(2j+1)!}$$

For any fixed n, the first factor on the right is just a constant, and the second factor is just one over a polynomial of degree $(n-3)/2$ in the index j. Therefore, after some number of terms, the series solution goes over to a simple factorial form with a polynomial divisor. It can be shown that the resulting function $f(r)$ is such that Huygens' principle is satisfied, so this implies that the principle is satisfied for any odd number of space dimensions.

In contrast, if the number of space dimensions is *even,* we do not have interleaving of the factors in the denominator of the coefficients. In this case we can only re-write in the form

$$c_{2j} = c_0 \frac{\left(\frac{n}{2}-1\right)!}{2^{2j}\,j!\left(\frac{n}{2}-1+j\right)!}(-1)^j$$

For example, in the case $n = 2$ (i.e., two spatial dimensions) we have the coefficients

$$c_{2j} = c_0 \frac{(-1/4)^j}{(j!)^2}$$

This gives the function

$$f(r) = c_0\left[1-\frac{1}{(1!)^2}\left(\frac{r^2}{4}\right)+\frac{1}{(2!)}\left(\frac{r^2}{4}\right)^2-\frac{1}{(3!)^2}\left(\frac{r^2}{4}\right)^3+\ldots\right]$$

This is the Bessel function of order zero, often denoted as J_0. A plot of this function is shown below.

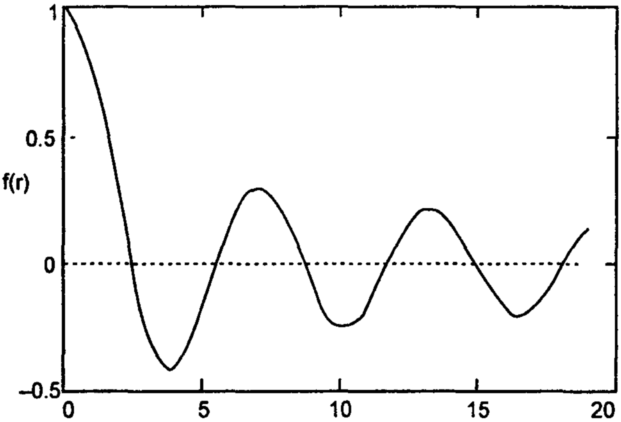


Fig. 12 Bessel Function

For positive arguments *r*, the Bessel function $J_0(r)$ can be expressed as

$$J_0(r) = \frac{2}{\pi}\int_0^{\infty}\sin\left(\cosh(\theta)r\right)d\theta$$

Multiplying through by the temporal solution *g*(*t*) = *sin*(*ct*) gives

$$\psi(r,t) \;=\; \frac{1}{\pi}\int_0^{\infty}\left[\cos\left(\cosh(\theta)r - ct\right) - \cos\left(\cosh(\theta)r + ct\right)\right]d\theta$$

Hence, instead of the solution being purely a function of *r* ± *ct* as in the case of odd dimensions, we find that it is an integral of functions of *cosh*(*θ*)*r* ± *ct*. Each value of *θ* corresponds to a propagation speed of *c*/*cosh*(*θ*), so the speeds vary from *c* down to zero. This signifies that the wave function at any event is correlated not just with the wave function on its 'light cone', but with the wave function at every event inside its light cone.

It would be interesting to work out the connections between Huygens' principle and the zeta function (whose value can only be given in simple closed form for even arguments) and the Bernoulli numbers (which are non-zero only for even indices). It's also interesting to note the analogy between Huygens' spherical wavelets centred on the boundary of the wave front and the technique of analytic continuation, by which we expand the boundary of an analytic region by means of disks of convergence centred on or near the boundary of the existing analytic region.

Paul Dirac gave an interesting general argument for a much stronger version of Huygens' principle in the context of quantum mechanics. In his *Principles of Quantum Mechanics* he noted that a measurement of a component of the instantaneous velocity of a free electron must give the value ±*c*, which implies that electrons (and massive particles in general) always propagate along null intervals, i.e., on the local light cone. At first this may seem to contradict the fact that we observe massive objects to move at speeds much less than the speed of light, but Dirac points out that observed velocities are always average velocities over appreciable time intervals, whereas the equations of motion of the particle show that its velocity oscillates between +*c* and –*c* in such a way that the mean value agrees with the average value. He argues that this must be the case in any relativistic theory that incorporates the uncertainty principle, because in order to measure the velocity of a particle we must measure its position at two different times, and then divide the change in position by the elapsed time. To approximate as closely as possible to the instantaneous velocity, the time interval must go to zero, which implies that the position measurements must approach infinite precision. However, according to the uncertainty principle, the extreme precision of the position measurement implies an approach to infinite indeterminacy in the momentum, which means that almost all values of momentum—from zero to infinity—become equally probable. Hence the momentum is almost certainly infinite, which corresponds to a speed of ±*c*. This is obviously a very general argument, and applies to all massive particles.

According to Newton's laws, the incremental work *dW* done by a force f on a particle moving an incremental distance *dx*, *dy*, *dz* in 3-dimensional space is given by the dot product.

$$dW = f_x\,dx + f_y dy + f_z dz$$

Now suppose the particle is constrained in such a way that its position has only two degrees of freedom. In other words, there are two generalized position coordinates *X* and *Y* such that the position coordinates *x*, *y*, and *z* of the particle are each strictly functions of these two generalized coordinates. We can then define a generalized force *F* with the components F_X and F_Y such that

$$dW = F_x dX + F_Y dY$$

The total differentials of *x*, *y*, and *z* are then given by

$$dx \frac{\partial x}{\partial X} dX + \frac{\partial x}{\partial Y} dY \quad dy = \frac{\partial y}{\partial X} dX + \frac{\partial y}{\partial Y} \quad dz = \frac{\partial z}{\partial X} dX + \frac{\partial z}{\partial Y} dY$$

Substituting these differentials into and collecting terms by *dX* and *dY*, we have

$$dW = \left(f_x \frac{\partial x}{\partial X} + f_y \frac{\partial y}{\partial X} + f_z \frac{\partial z}{\partial X} \right) dX + \left(f_x \frac{\partial x}{\partial Y} + f_y \frac{\partial y}{\partial Y} + f_z \frac{\partial z}{\partial Y} \right) dY$$

Comparing this with Equation, we see that the generalized force components are given by

$$F_X = f_x \frac{\partial x}{\partial X} + f_y \frac{\partial y}{\partial X} + f_z \frac{\partial z}{\partial X} \quad F_Y = f_x \frac{\partial x}{\partial Y} + f_y \frac{\partial y}{\partial Y} + f_z \frac{\partial z}{\partial Y}$$

Now, according to Newton's second law of motion, the individual components of force for a particle of mass m are

$$f_x = m\frac{d\dot{x}}{dt},\quad f_y = m\frac{d\dot{y}}{dt},\quad f_z = m\frac{d\dot{z}}{dt}$$

Substituting into the expression for F_X gives

$$F_x \quad = m\left(\frac{d\dot{x}}{dt}\frac{\partial x}{\partial X} + \frac{d\dot{y}}{dt}\frac{\partial y}{\partial X} + \frac{d\dot{z}}{dt}\frac{\partial z}{\partial X}\right)$$

and similarly for F_Y. Notice that the first product on the right side can be expanded as

$$\frac{d\dot{x}}{dt}\frac{\partial x}{\partial X} = \frac{d}{dt}\left(\dot{x}\frac{\partial x}{\partial X}\right) - \dot{x}\frac{d}{dt}\left(\frac{\partial x}{\partial X}\right)$$

and similarly for the other two products. Since x and X are both strictly functions of t, it follows that partial differentiation with respect to t is the same as total differentiation, and so the order of differentiation in the right-most term of can be reversed (because partial differentiation is commutative). Hence Equation can be written as

$$\frac{d\dot{x}}{dt}\frac{\partial x}{\partial X} = \frac{d}{dt}\left(\dot{x}\frac{\partial}{\partial X}\right) - \dot{x}\frac{\partial}{\partial X}\left(\frac{dx}{dt}\right)$$

Substituting this (and the corresponding expressions for the other two products) into equation, we get

$$\frac{F_X}{m} = \frac{d}{dt}\left(\dot{x}\frac{\partial x}{\partial X} + \dot{y}\frac{\partial y}{\partial X} + \dot{z}\frac{\partial z}{\partial X}\right) - \left(\dot{x}\frac{\partial \dot{x}}{\partial X} + \dot{y}\frac{\partial \dot{y}}{\partial X} + \dot{z}\frac{\partial \dot{z}}{\partial X}\right)$$

Variations in x,y,z and X at constant t are independent of t (since each of these variables is strictly a function of t), so we have

$$\frac{\partial x}{\partial X} = \frac{\partial \dot{x}}{\partial \dot{X}} \quad \frac{\partial y}{\partial X} = \frac{\partial \dot{y}}{\partial \dot{X}} \quad \frac{\partial z}{\partial X} = \frac{\partial \dot{z}}{\partial \dot{X}}$$

Making these substitutions into gives

$$\frac{F_X}{m} = \frac{d}{dt}\left(\dot{x}\frac{\partial \dot{x}}{\partial \dot{X}} + \dot{y}\frac{\partial \dot{y}}{\partial \dot{X}} + \dot{z}\frac{\partial \dot{z}}{\partial \dot{X}}\right) - \left(\dot{x}\frac{\partial \dot{x}}{\partial X} + \dot{y}\frac{\partial \dot{y}}{\partial X} + \dot{z}\frac{\partial \dot{z}}{\partial X}\right)$$

Each term now contains an expression of the form $r(\partial r/\partial s)$, which can also be written as $\partial(r^2/2)/\partial s$, so the overall expression can be re-written as

$$F_\lambda = \frac{d}{dt}\left(\frac{\partial}{\partial \dot{X}}\left[m\frac{\dot{x}^2\dot{y}^2+\dot{z}^2}{2}\right]\right)-\frac{\partial}{\partial X}\left(\left[m\frac{\dot{x}^2+\dot{y}^2+\dot{z}^2}{2}\right]\right)$$

The quantity inside the square brackets is simply the kinetic energy, conventionally denoted by T. Thus the generalized force F_X, and similarly the generalized force F_Y, can be expressed as

$$F_X = \frac{d}{dt}\left(\frac{\partial T}{\partial \dot{X}}\right)-\frac{\partial T}{\partial X} \quad F_Y = \frac{d}{dt}\left(\frac{\partial T}{\partial \dot{Y}}\right)-\frac{\partial T}{\partial Y}$$

These are the Euler-Lagrange equations of motion, which are equivalent to Newton's laws of motion. (Notice that if X is identified with x in equation, then F_X reduces to Newton's expression for f_x, and likewise for the other components.)

If the total energy is conserved, then the work done on the particle must be converted to potential energy, conventionally denoted by V, which must be purely a function of the spatial coordinates x,y,z, or equivalently a function of the generalized configuration coordinates X,Y, and possibly the derivatives of these coordinates, but independent of the time t. (The independence of the Lagrangian with respect to the time coordinate for a process in which energy is conserved is an example of Noether's theorem, which asserts that any conserved quantity, such as energy, corresponds to a symmetry, i.e., the independence of a system with respect to a particular variable, such as time.) If the potential depends on the derivatives of the position coordinates it is said to be a velocity-dependent potential, as discussed in the note on Gerber's Gravity. However, most potentials depend only on the position coordinates and not on their derivatives. In that case we have

$$dW = -dV = -\frac{\partial V}{\partial X}dX - \frac{\partial V}{\partial Y}dY$$

Comparing this with equation, we see that

$$F_X = -\frac{\partial V}{\partial X} \quad F_Y = -\frac{\partial V}{\partial Y}$$

and therefore the Euler-Lagrange equations for conservative systems can be written as

$$-\frac{\partial V}{\partial Y} = \frac{d}{dt}\left(\frac{\partial T}{\partial \dot{X}}\right)-\frac{\partial T}{\partial X} \quad -\frac{\partial V}{\partial Y} = \frac{d}{dt}\left(\frac{\partial T}{\partial \dot{Y}}\right)-\frac{\partial T}{\partial Y}$$

Rearranging terms, we have

$$\frac{\partial(T-V)}{\partial X}=\frac{d}{dt}\left(\frac{\partial T}{\partial \dot{X}}\right) \quad \frac{\partial(T-V)}{\partial Y}=\frac{d}{dt}\left(\frac{\partial T}{\partial \dot{Y}}\right)$$

Furthermore, since *V* is purely a function of the configuration variables, independent of their rates of change, we can just as well substitute (*T–V*) in place of *T* on the right sides of these equations, so in terms of the parameter $L = T - V$ these equations can be written simply as

$$\frac{\partial L}{\partial X}=\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{X}}\right) \quad \frac{\partial L}{\partial Y}=\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{Y}}\right)$$

The quantity *L* is called the Lagrangian. This derivation was carried out for a single particle moving with two degrees of freedom in three-dimensional space, but the same derivation can be applied to collections of any number of particles. For a set of *N* particles there are *3N* configuration coordinates, but the degrees of freedom will often be much less, especially if the particles form rigid bodies. Letting q_1, q_2, .., q_n denote a set of generalized configuration coordinates for a conservative physical system with n degrees of freedom, the equations of motion of the system are

$$\frac{\partial L}{\partial q_j}=\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}_j}\right) \quad j = 1, 2, \ldots\ldots, n$$

where *L* is the Lagrangian of the system, i.e., the difference between the kinetic and the potential energies, expressed in terms of the generalized coordinates and their time derivatives. These equations are usually credited jointly to Euler along with Lagrange, because although Lagrange was the first to formulate them specifically as the equations of motion, they were previously derived by Euler as the conditions under which a point passes from one specified place and time to another in such a way that the integral of a given function *L* with respect to time is stationary. This is a fundamental result in the calculus of variations, and can be applied to fairly arbitrary functions *L* .

To illustrate the application of these equations, consider a simple mass-spring system, consisting of a particle of mass *m* on the *x* axis attached to the end of a massless spring with spring constant *k* and null point at $x = 0$. For any position x, the spring exerts a force equal to $F = kx$, and the potential energy is the integral of force with respect to displacement. Similarly the kinetic energy is the integral of the inertial force $F = ma$ with respect to displacement. Thus the kinetic and potential energies of the system are

$$T = \int m\frac{dv}{dt}dx = m\int vdv = \frac{1}{2}mv^2 \quad V = \int kxdx = \frac{1}{2}kx^2$$

Therefore the Lagrangian of the system is

$$L(x, x) = \frac{1}{2}m\dot{x}^2 - \frac{1}{2}kx^2$$

The partial derivatives are

$$\frac{\partial L}{\partial x} = -kx \quad \frac{\partial L}{\partial \dot{x}} = m\dot{x}$$

Substituting into Lagrange's equation, we get the familiar equation of harmonic motion for a mass-spring system

$$-kx = \frac{d}{dt}\left(m\dot{x}\right) = m\ddot{x}$$

Of course, this simply expresses Newton's second law, $F = ma$, for the particle. It's also equivalent to the fact that the total energy $E = T + V$ is constant, as can be seen by differentiating E with respect to t and then dividing through by dx/dt.

The equivalence between the Lagrangian equation of motion (for conservative systems) and the conservation of energy is a general consequence of the fact that the kinetic energy of a particle is strictly proportional to the square of the particle's velocity. Of course, in terms of the generalized parameters, it's possible for the kinetic energy to be a function of both q and $\dot{q}$, but since the transformation $dx = (\partial X/\partial q)dq$ between x and q is equivalent to $dx/dt = (\partial x/\partial q)\, dq/dt$, it follows that for a fixed configuration the kinetic energy is proportional to the squares of the generalized velocity parameters. Therefore, in general, we have

$$\frac{\partial T}{\partial \dot{q}}\dot{q} = 2T = \frac{\partial L}{\partial \dot{q}}\dot{q}$$

where we've made use of the fact that the potential energy V (for conservative systems) is independent of $\dot{q}$. Now, the total energy is

$E = T + V = 2T - L$, so the conservation of energy can be expressed in the form

$$\frac{d(2T-L)}{dt} = \frac{d(2T)}{dt} - \frac{dL}{dt} = 0$$

The two terms on the right hand side can be expanded as

$$\frac{d(2T)}{dt} = \frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}}\dot{q}\right) = \frac{\partial L}{\partial \dot{q}}\ddot{q} + \dot{q}\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}}\right)$$

$$\frac{dL}{dt} = \frac{\partial L}{\partial q}\frac{dq}{dt} + \frac{\partial L}{\partial \dot{q}}\frac{d\dot{q}}{dt}$$

Substituting into the previous equation and dividing through by $\dot{q}$ (applying analytic continuation to remove the singularity when $\dot{q} = 0$), we see that the conservation of energy implies

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}}\right) - \frac{\partial L}{\partial q} = 0$$

which is just Lagrange's equation of motion. Of course, the same derivation applies to any number of particles, and their generalized coordinates.

The correspondence between the conservation of energy and the Lagrangian equations of motion suggests that there might be a convenient variational formulation of mechanics in terms of the total energy $E = T + V$ (as opposed to the Lagrangian $L = T-V$). Notice that the partial derivative of L with respect to x' is the momentum of the particle. In general, given the Lagrangian, we can define the generalized momenta as

$$p_j = \frac{\partial L}{\partial \dot{q}_j} = \frac{\partial T}{\partial \dot{q}_j} = \frac{\partial (T+V)}{\partial \dot{q}_j}$$

(The partial of V is zero, so it's inclusion and sign in this definition is a matter of convention.) Thus to each generalized configuration coordinate q_j there corresponds a generalized momenta p_j. In our simple mass-spring example with the single generalized coordinate $q = x$, the total energy $H = T + V$ in terms of these conjugate parameters is

$$H(q, p) = \frac{p^2}{2m} + \frac{1}{2}kq^2$$

The function *H(q, p)* is called the Hamiltonian of the system. Taking the partial derivatives of H with respect to *p* and *q*, we have

$$\frac{\partial H}{\partial p} = \frac{p}{m} \quad \frac{\partial H}{\partial q} = kq$$

Notice that, in this example, *p/m* equals *q'* (essentially by definition, since *p* = *mv*), and *kq* equals + *p'* (by the equation of motion). In general it can be shown that, for any conservative system with generalized coordinates q_j and the corresponding momenta p_j, if we express the total energy *H* in terms of the q_j and p_j, then we have

$$\frac{\partial H}{\partial p_j} = \dot{q}_j \quad \frac{\partial H}{\partial q_j} = -\dot{p}_j$$

These are Hamilton's equations of motion. Although they are strictly equivalent to Lagrange's and Newton's equations, the equations of Hamilton have proven to be more suitable for adaptation to quantum mechanics. The Lagrangian and Hamiltonian formulations of mechanics are also notable for the fact that they express the laws of mechanics without reference to any particular coordinate system for the configuration space. Of course, in their original forms, they assumed an absolute time coordinate and perfectly rigid bodies, but with suitable restrictions they can be adapted to relativistic mechanics as well.

In quantum mechanics, a pair of conjugate variables q_j, p_j, such as position and momentum, generally do not commute, which means that the operation consisting of a measurement of q_j followed by a measurement of p_j is different than the operation of performing these measurements in the reverse order. This is because the eigenstates corresponding to the respective measurement operators are incompatible. As a result, the system cannot simultaneously have both a definite value of q_j and a definite value of p_j.

Fourier Transforms and Uncertainty

The function *exp*($-x^2$) has no simple closed-form indefinite integral, but the related function *x exp*($-x^2$) does have a simple integral, namely,

$$\int xe^{-x^2} = -\frac{1}{2}e^{-x^2}$$

This identity can be used to evaluate the definite integral of *exp*($-x^2$) from $x = -\infty$ to $+\infty$. Letting Q denote the value of this definite integral, we can write

$$Q^2 = \left(\int_{-\infty}^{\infty} e^{-x^2} dx\right)\left(\int_{-\infty}^{\infty} e^{-y^2} dy\right) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\left(x^2+y^2\right)} dxdy$$

In terms of polar coordinates on the x, y plane we have $x = r$ $cos(\theta)$ and $y = r\ sin(\theta)$, and therefore $x^2 + y^2 = r^2$. The Jacobian of the transformation is

$$\begin{vmatrix} \partial x/\partial r & \partial x/\partial\theta \\ \partial y/\partial r & \partial y/\partial\theta \end{vmatrix} = \begin{vmatrix} \cos(\theta) & -r\sin(\theta) \\ \sin(\theta) & r\cos(\theta) \end{vmatrix} = r$$

so the incremental area element is

$$dA = dxdy = r\, dr\, d\theta$$

hence the double integral expressed in terms of r, q coordinates is

$$Q^2 = \int_0^{2\pi}\int_0^{\infty} e^{-r^2} r\, dr\, d\theta$$

Making use of equation to evaluate the interior definite integral, we have

$$Q^2 = \int_0^{2\pi}\left(\int_0^{\infty} e^{-r^2} r\, dr\right) d\theta = \int_0^{2\pi} \frac{1}{2} d\theta = \pi$$

Therefore the definite integral of *exp*($-x^2$) from $-\infty$ to ∞ is

$$Q = \sqrt{\pi} = \int_{-\infty}^{\infty} e^{-x^2} dx$$

More generally, if we replace the exponent $-x^2$ with $-ax^2 + bx + c$, we can define the parameter

$$y = x\sqrt{a} - \frac{b}{2\sqrt{a}}$$

in terms of which we can write

$$\int_{-\infty}^{\infty} e^{-ax^2+bc+c} dx = \frac{e^{\left(b^2-4ac\right)/(4a)}}{\sqrt{a}} \int_{-\infty}^{\infty} e^{-y^2} dy = e^{\left(b^2-4ac\right)/(4a)}\sqrt{\frac{\pi}{a}}$$

This definite integral is particularly useful when considering the

Fourier transform of a normal density distribution. Recall that for any function $f(x)$ we can define another function $F(y)$ that satisfies the relations

$$F(y) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(x)^{ixy}\, dx$$

$$F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} F(y) e^{-ixy} dy$$

These two functions are a *Fourier transform pair*, i.e., each of them is the Fourier transform of the other. Now, if $f(x)$ is the normal probability density function

$$f(x) = \frac{e - [x-\mu]^2 2\sigma^2}{\sigma\sqrt{2\pi}}$$

then the Fourier transform is

$$F(y) = \frac{1}{2\pi\sigma} \int_{-\infty}^{\infty} e^{-[x-\mu]^2/2\sigma^2+ixy} dx$$

so the exponent in the integral is of the form $-ax^2 + bx + c$ with $a = 1/2\sigma^2$, $b = iy + \mu/\sigma^2$, and $c = -\mu^2/(2\sigma^2)$. Hence the Fourier transform of the normal density distribution is

$$\mathrm{F(y)} = e^{\mu^2/(2\sigma^2)} \frac{e^{-(y-i\mu/\sigma^2)^2/(2/\sigma^2)}}{\frac{1}{\sigma}\sqrt{2\pi}}$$

Choosing our scales so that the mean of f is zero, i.e., so that $\mu = 0$, the above expression reduces to

$$F(y) = \frac{e^{-y^2/(2/\sigma^2)}}{\frac{1}{\sigma}\sqrt{2\pi}}$$

In other words, the Fourier transform of the normal distribution with mean zero and standard deviation σ is also a normal distribution with mean zero, but with standard deviation $1/\sigma$. This shows that the variances of the f and F distributions satisfy the 'uncertainty relation' $var(f)\ var(F) = 1$. This equality is the limiting case of a general inequality on the product of variances of Fourier transform pairs. In general, if f is an arbitrary probability density distribution and F is its Fourier transform, then

$$var\,(f)\;var\,(F) \geq$$

Notice that *f* and *F* are just two different ways of characterizing the same distribution, one in the amplitude domain and the other in the frequency domain. Given either of these distributions, the other is completely determined.

The relation between conjugate variables (such as position and momentum) in quantum mechanics can be expressed in terms of the relation between Fourier transform pairs. Consider a physical system with just one degree of freedom, represented by the operator q, and let p denote the operator for the corresponding momentum (i.e., the generalized momentum of q in the usual Hamiltonian formulation). The basic commutation relation between these operators is

$$qp - pq = \hbar$$

Notice the symmetry between the p and q operators if we replace i with $-i$. The usual Schrodinger representation of this system takes the observable q as a 'diagonal' operator, i.e., with the eigenvalues specified explicitly, and then the corresponding momentum operator is defined as

$$p = i\hbar\frac{d}{dq}$$

However, we could (in theory) just as well take p as a diagonal operator, and then q would be given by

$$q = i\hbar\frac{d}{dp}$$

Dirac referred to this as the momentum representation of the system. This again shows the symmetry between q and p under an exchange of i and $-i$. Indeed if we let $<q|S>$ and $<p|S>$ for any given state S denote the probability amplitudes that measurements corresponding to the operators q and p will return the eigenvalues q and p respectively, then we find

$$\langle p|S\rangle = \frac{1}{\hbar}\int_{-\infty}^{\infty}\langle q|S\rangle e^{-iqp/\hbar}dq$$

$$\langle q|S\rangle = \frac{1}{\hbar}\int_{-\infty}^{\infty}\langle p|S\rangle e^{-iqp/\hbar}dp$$

In other words, the probability amplitude distributions of two conjugate variables are simply the (suitably scaled) Fourier transforms of each other. We saw previously that the dispersions (variances) of two density distributions that comprise a Fourier transform pair satisfy the inequality (2), so the variances of the probability amplitude distributions of conjugate observables in quantum mechanics satisfy such an inequality. Thus Heisenberg's uncertainty principle for conjugate pairs of observables follows directly from the fact that those observables are essentially the Fourier transforms of each other.

Of course, this attribute of Fourier transform pairs is purely mathematical, and has no *apriori* applicability to pairs of observables such as position and momentum, or time and energy. The physical content of quantum mechanics is based on the two relations

$$E = \hbar\omega \quad p = \hbar k$$

where E is energy, p is momentum (in one dimension),$\hbar$ is Planck's (reduced) constant, ω is the frequency with units second^{-1}, and k is the wave number with units metre^{-1}. These relations were introduced in the early 1900s by Planck, Einstein, and Broglie to account for non-classical phenomena such as cavity radiation and the photoelectric effect, both of which depend on the particle-like behaviour of entities that had previously been modeled as waves, as well as phenomena involving wave-like behaviour of material particles. These are the relations that associate the familiar observables of energy, momentum, space, and time, with the frequency domain. Indeed in terms of the characteristic time $\tau = 1/\omega$ and distance $\lambda\!\!\!^{-} = 1/k$ the above relations can be written as

$$\tau E = \lambda\!\!\!^{-} p = \hbar$$

which already clearly reveals the conjugacy of time and energy, and of distance and momentum. In view of this, it isn't surprising to find that the product of the dispersions of two conjugate observables (such as position and momentum) cannot be less than one quanta of action, represented by $\hbar$.

In a sense, there is also a conjugacy between space and time—two observable that had been regarded as disjoint and independent prior to the early 1900s. In special relativity the inertial space and time intervals *dx* and *dt* between two events are components of a single invariant spacetime interval ds between those events. These intervals are related according to the Minkowski metric, which can be written in the form

$$\left(\frac{dx}{dt}+\frac{ds}{dt}\right)\left(\frac{dx}{dt}-\frac{ds}{dt}\right)=\frac{1}{c^2}$$

This can be regarded as an 'uncertainty relation' for space and time. In general, physics was based, prior to 1900, on the premise that $\hbar$ and $1/c^2$ were both zero. With the advent of quantum mechanics and special relativity, it was realised that they both have non-zero values, although they are extremely small in terms of ordinary units.

Spherical Waves in Higher Dimensions

As discussed in the section on Huygens' principle, if we separate the solution $y(r,t)$ of the usual wave equation in n-dimensional space (with one time dimension) into a time component and a spatial component, we have $y(r,t) = f(r)g(t)$, and the spatial and temporal components satisfy the individual equations

$$r\frac{d^2f}{dr^2}+(n-1)\frac{df}{dr}+k^2rf=0 \quad \frac{d^2g}{dt^2}+\omega^2g=0$$

where k is a constant with units of 1/distance and ω is a constant with units of 1/time. Thus the temporal component satisfies the simple harmonic equation, with a general solution of the form $g(t) = g_1 e^{i\omega t} + g_2 e^{-i\omega t}$ where g_1 and g_2 are arbitrary constants. In just one spatial dimension ($n = 1$) the spatial equation also reduces to the simple harmonic equation, with the general solution $f(r) = f_1 e^{ikr} + f_2 e^{-ikr}$ for constants f_1 and f_2. Combining these, we get the wave function

$$\psi(r,t)=\left(f_1e^{ikr}+f_2e^{-ikr}\right)\left(g_1{}^{i\omega t}+g_2e^{-i\omega t}\right)$$

$$=f_1g_1e^{i(kr+\omega t)}+f_1g_2e^{i(kr-\omega t)}+f_2g_1e^{-i(kr-\omega t)}+f_2g_2e^{-i(kr+\omega t)}$$

Thus the solution is a sum of functions of the quantities $kr + \omega t$ and $kr - \omega t$. If we require that f and g are real-valued, then $g(t) = g_1 \cos(wt)$, $f(r) = f_1 \cos(kr)$, and

$$\psi(r,t) = 2f_1g_1[\cos(kr + \omega t) + \cos(kr - \omega t)]$$

More generally, we can verify by direct substitution that the one-dimensional wave equation is satisfied by any function of the form

$$\psi(r, t) = A(\omega t - kr) + B(\omega t + kr)$$

where $A(x)$ and $B(x)$ are quite arbitrary functions. In effect, Huygens' principle can be read directly from this equation, since it implies that a pulse disturbance propagates sharply at a constant speed. We also know that with three spatial dimensions ($n = 3$) the general spatial solution is $f(r) = cos(kr)/r$, and again Huygens' principle applies. We asserted that a similar result obtains for any odd number of spatial dimensions. To show this more explicitly, recall that spatial equation is

$$r\frac{d^2 f}{dr^2}+(n-1)\frac{df}{dr}+k^2 rf=0$$

We assume $f(r)$ has a power series expansion

$$f(r) = c_0 r^q + c_1 r^{1+q} + c_2 r^{2+q} + c_3 r^3 + q + \ldots$$

where c_0 is the first non-zero coefficient and q is non-negative (to ensure that $f(r)$ is finite at $r = 0$). Inserting this series and its derivatives into equation and setting the coefficient of each power of r to zero, we get the conditions

$$c_0 q(q - 2 + n) = 0$$
$$c_1(q + 1)(q - 1 + n) = 0$$
$$c_2(q + 2)(q + n) + k^2 c_0 = 0$$
$$c_3(q + 3)(q + n + 1) + k^2 c_1 = 0$$
$$c_4(q + 4)(q + n + 2) + k^2 c_2 = 0$$

and so on. Since c_0 is stipulated to be non-zero, and since we are requiring q to be non-negative, the first of these conditions implies $q = 0$, and so the second implies $c_1 = 0$. The remaining conditions give $c_j + 2$ as a multiple of c_j, so it follows that $c_j = 0$ for all odd j. The coefficients with even indices are then given by

$$c_2=\frac{-k^2 c_0}{(2)(n)} \quad c_4=\frac{-k^2 c_2}{(4)(n+2)} \quad c_6=\frac{-k^2 c_4}{(6)(n+4)} \quad \ldots$$

With $n = 3$ the spatial component of the wave function is therefore

$$f_3^{(r)}=c_0\left[1+\frac{(-k^2)}{3!}r^2+\frac{(-k^2)}{5!}r^4+\frac{(-k^2)}{7!}r^6+\ldots\right]$$

$$= \frac{c_0}{kr}\left[\frac{kr}{1!} - \frac{(kr)^3}{3!} + \frac{(kr)^5}{5!} - \frac{(kr)^7}{7!} + \dots\right]$$

$$= \frac{c_0}{kr}\sin(kr)$$

On the other hand, in a five-dimensional space, we have *n* = *5*, and the spatial part of the wave function is

$$f_5(r) = c_0\left[1 + \frac{(-k^2)}{(2.5)}r^2 + \frac{(-k^2)^2}{(2.5)(4.7)}r^4 + \frac{(-k^2)^2}{(2.5)(4.7)(6.9)}r^6 + \dots\right]$$

$$= \frac{3c_0}{(kr)^3}\left[\frac{(kr)^3}{3.1!} - \frac{(kr)^5}{5.3!} + \frac{(kr)^7}{7.5!} - \frac{(kr)^9}{9.7!} + \dots\right]$$

Putting s = kr and letting $\alpha(s)$ denote the expression inside the last square brackets, we see that

$$\frac{1}{s}\frac{d\alpha}{ds} = \sin(s)$$

Multiplying through by *s* and integrating, we find that $\alpha(s) = sin(s) - s\ cos(s)$, so we have

$$f_5(r) = \frac{3c_0}{(kr)^3}\left[\sin(kr) - kr\cos(kr)\right]$$

The same general approach allows us to determine the closed-form expression for $f_n(r)$ for any odd number of dimensions *n*. For example, in seven spatial dimensions (*n* = *7*) we have the series

$$f_7(r) = c_0\left[1 + \frac{(-K^2)}{(2\cdot 7)}r^2 + \frac{(-K^2)^2}{(2{:}7)(4\cdot 9)}r^4\right.$$

$$\left. + \frac{(-K^2)^2}{(2\cdot 7)(4\cdot 9)(6\cdot 11)}r^6 + \dots\right]$$

$$= \frac{15c_0}{(kr)^5}\left[\frac{(kr)^5}{(3\cdot 5)\cdot 1!} - \frac{(kr)^7}{(5\cdot 7)\cdot 3!} + \frac{(kr)^9}{(7\cdot 9)\cdot 5!} - \frac{(kr)^{11}}{(9\cdot 11)\cdot 7!} + ..\right]$$

Again putting $s = kr$ and letting $\alpha(s)$ denote the expression inside the last square brackets, we see that

$$\frac{1}{s}\frac{d}{ds}\left(\frac{1}{s}\frac{d\alpha}{ds}\right) = \sin(s)$$

Multiplying and integrating twice gives

$$\alpha(s) = 3\sin(s) - 3s\cos(s) - s^2\sin(s)$$

Therefore, the spatial part of the spherical solution of the wave equation in seven space dimensions is

$$f_7(r) = \frac{15c_0}{(kr)^5}\left[3\sin(kr) - 3kr\cos(kr) - (kr)^2\sin(kr)\right]$$

The same approach leads to the solution in nine space dimensions

$$f_9(r) = \frac{105c_0}{(kr)^7}\Big[15\sin(kr) - 15kr\cos(kr) - 6(kr)^2$$

$$\sin(kr) + (kr)^3\cos(kr)\Big]$$

and so on. In general, the spatial part of the spherical wave solution in n dimensions has the form

$$f_n(r) = \frac{K_n c_0}{(kr)^{n-2}}\left[A_n(kr)\sin(kr) - kr\,B_n(kr)\cos(kr)\right]$$

where the constant K_n equals $(n-2)(n-4)\ldots$, and the expressions $A_n(kr)$ and $B_n(kr)$ denote ‘even’ polynomials in kr. These polynomials for the first several odd values of n are listed below.

$$\begin{array}{llllll}
A_1(s) & = & 0 & B_1(s) & = & -1 \\
A_3(s) & = & 1 & B_3(s) & = & 0 \\
A_5(s) & = & 1 & B_5(s) & = & 1 \\
A_7(s) & = & 3-s^2 & B_7(s) & = & 3 \\
A_9(s) & = & 15-6s^2 & B_9(s) & = & 15-s^2 \\
A_{11}(s) & = & 105-45s^2+s^4 & B_{11}(s) & = & 105-45s^2+s^4 \\
A_{13}(s) & = & 945-420s^2+15s^4 & B_{13}(s) & = & 945-105s^2+s^4
\end{array}$$

In general, for odd n greater than 1, these polynomials can be expressed as

$$A_{2m+3}(s) = \sum_{j=0}^{m/2}(-1)^{j}\frac{2(m-j)}{2^{m-j}j!(m-j)!}s^{2j}$$

$$B_{2m+3}(s) = \sum_{j=0}^{(m-1)/2}(-1)^{j}\frac{2(m-j)-1]!}{2^{m-1-j}j!(m-1-j)!}s^{2j}$$

Interestingly, referring to the article on proving that π is irrational, we see that the spatial functions $f_n(r)$ for odd integers n greater than 1 can also be expressed (up to a constant factor) as a simple integral

$$f_{2k+3} = \int_{-1}^{+1}\left(1-x^2\right)^k \cos(rx)\,dx$$

The figure 13 shows $f_n(r)$ for spaces of one, three, and five dimensions.

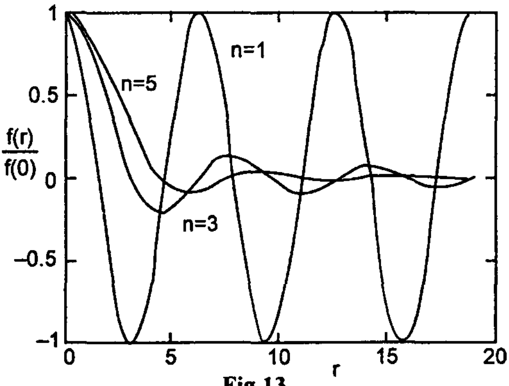


Fig.13

We can use linear combinations of solutions of this form to generate arbitrary waveforms.

As an aside, although the case of negative n presumably has no physical significance (negative dimensions?), we note that a similar approach enables us to solve equation for these cases as well. If n is even, the coefficients are undefined (because they involve a division by zero), but we get well-defined functions for odd n. The solutions for the first couple of odd negative values of n are

$$f_{-1(r)} = c_0[1-\cos(kr)-kr\sin(kr)]$$

$$f_{-3(r)} = c_0\left[1+\frac{k^2}{6}r^2+\frac{1}{3}\left(3\cos(kr)+3kr\sin(kr)\right)-(kr)^2\cos(kr)\right]$$

These solutions increase to infinity as r increases, unlike the solutions for positive n, which drop to zero as r increases.

Returning to positive dimensions, there's an interesting relationship between the 'basis functions' of the general spherical wave solutions in successive odd dimensions. The wave equation in spherical coordinates in n-dimensional space can be written in terms of the differential operator.

$$\Diamond_n = \frac{\partial^2}{\partial r^2} + \frac{n-1}{r}\frac{\partial}{\partial r} - \frac{\partial^2}{\partial t^2}$$

A function $\psi(r,t)$ is a solution of the n-dimensional wave equation if and only if

$$\Diamond_n \psi(r,t) = 0$$

Knowing the general form of the spatial part $f_n(r)$, and the simple temporal part for all n, we can write a basic combined space-time solution for odd n as

$$\phi_n(r,t) = \frac{1}{r^{n-2}}\left[A_n(r)\sin(r \pm t) - rB_n(r)\cos(r \pm t)\right]$$

where $A_n(r)$ and $B_n(r)$ are the polynomials defined previously. By direct substitution it can be verified that successive odd basis solutions are related by

$$\Diamond_m f_n = (n-m)\phi_{n+2}$$

This is a remarkable fact, signifying that a basis solution for n space dimensions not only satisfies the wave equation everywhere in n-dimensional space, it also satisfies the wave equation in spaces of every odd number of dimensions at every radius and time for which the basis solution for $n+2$ dimensions vanishes. Hence there is an infinite sequence of expanding (or contracting) discrete shells on which any basis solution satisfies the wave equation for spaces of all odd dimensions.

In the special case $m = 1$ the spherical wave operator reduces to

$$\Diamond_1 = \frac{\partial^2}{\partial r^2} - \frac{\partial^2}{\partial t^2}$$

and we have the recurrence

$$f_{n+2} = \left(\frac{1}{n-1}\Diamond_1\right)\phi_n$$

We refer to the ϕ_n as *basis* solutions, because it's easy to see that if $\phi_n(r,t)$ is a solution, then so is $k\phi_n(jr, jt)$ for any constants k and j. For example, each of the following expressions are solutions for $n = 5$:

$$\frac{\sin(r\pm t)}{r^2}-\frac{\cos(r\pm t)}{r^3},\ \frac{\sin(2[\mathrm{r}\pm \mathrm{t}])}{(2r)^2}$$

$$-\frac{\cos(2[r\pm t])}{(2r)^2},\ \frac{\sin(3[r\pm t])}{(3r)^2}-\frac{\cos(3[r\pm t])}{(3r)^3}$$

and so on. Any linear combination of these solutions is also a solution, so by means analogous to Fourier series we can construct arbitrary functions of the form

$$F_s(r+t)=\sum_{k=1}^{\infty} c_k\cos(k(r+t))$$

$$F_s(r+t)=\sum_{k=1}^{\infty} c_k\cos(k(r+t))$$

and similarly for $G_s(r - t)$ and $G_c(r - t)$. The analogous functions can be constructed from the basis solutions for any odd n, so the spherical wave equation in n space dimensions is satisfied by

$$y(r,t)=\frac{A_n(r)F_s(r+t)-r^B{}_n(r)F_c(r+t)}{r^{n-2}}$$

$$+\frac{A_n(r)G_s(r-t)-r^B{}_n(r)G_c(r-t)}{r^{n-2}}\ .$$

for arbitrary functions F and G. Notice that the greater of the degree of A_n and the degree of rB_n is $(n - 3)/2$, so the lowest inverse power of r is $(n - 2) - (n - 3)/2 = (n - 1)/2$. This is consistent with the fact that the energy of a wave is proportional to the square of the amplitude, so the energy per unit 'area' of the spherical wave drops in inverse proportion to r^{n-1}, which is the dimension of the surface of a sphere in n-dimensional space.

It's interesting that, for odd space dimensions greater than 3, the amplitude contains formal terms that drop in inverse proportion to higher powers of r as well. This is in a sense misleading, because the amplitude (and hence the energy per unit area) is continuously changing as the wave propagates to greater values of r, and at the same time the value of the wave function is changing with phase. The wave function is not actually periodic, so the correspondence between energy and 'amplitude' contains an ambiguity, which manifests itself in the higher-order terms in higher dimensions.

Waves in Ascending and Descending Dimensions

There are some remarkable relationships between elementary spherically symmetrical solutions of the wave equation in spaces of different dimensions. As discussed in the note on spherical waves in higher dimensions, by the separation of variables we can split the wave equation into spatial and temporal parts, and the spatial part *f*(*r*) satisfies the differential equation.

$$\frac{d^2 f}{dr^2}+\frac{(n-1)}{r}\frac{df}{dr}+f=0$$

where *n* is the number of spatial dimensions. Letting f_n denote a solution of in a space of *n* dimensions (and one time dimension), the function given by

$$g(r)=-\frac{n}{r}\frac{df_n(r)}{dr}$$

is a solution of in space of *n* + *2* dimensions. To prove this, note that the definition of *g*(*r*) implies (after multiplying through by –*n*/*r* and differentiating)

$$\dot{f}_n=-\frac{r}{n}g \quad \ddot{f}_n=-\frac{1}{n}(r\dot{g}+g) \quad \dddot{f}_n=-\frac{1}{n}(r\ddot{g}+2\ddot{g})$$

where dots signify derivatives with respect to *r*. Also, differentiating equation gives

$$\dddot{f}_n+\frac{(n-1)r}{}\ddot{f}_n-\frac{(n-1)}{r^2}\dot{f}_n+\dot{f}_n=0$$

Substituting for the derivatives of *fn* from the preceeding expressions, we get

$$r\ddot{g}+2\dot{g}+\frac{(n-1)}{r}(r\dot{g}+g)-\frac{(n-1)}{r^2}rg+rg=0$$

Simplifying this expression, we get

$$\dot{f}_n=0$$

which proves that *g*(*r*) is a solution in space of *n* + *2* dimensions.

Thus we have a sequence of solutions in ascending dimensions

$$f_{n+2}(r) = -\frac{n}{r}\frac{df_n(r)}{dr}$$

Interestingly, this can also be expressed as a sequence of (scaled) solutions in descending dimensions. To see this, use equation to replace the left side of, and then multiply through by *–r/n* to give

$$\frac{1}{n}\left[r\ddot{f}_{n+2} + (n+1)\dot{f}_{n+2}\right] = \dot{f}_n$$

Integrating both sides, we get (up to a constant of integration)

$$\frac{1}{n}\left[r\dot{f}_{n+2} + nf_{n+2}\right] = f_n$$

Now we multiply through by r^{n-1} to give

$$\frac{1}{n}\left[r^n\dot{f}_{n+2} + nr^{n-1}f_{n+2}\right] = f_n r_{n-1}$$

The expression inside the square brackets is simply the derivative of $r^n f_{n+2}$, so this can be written as

$$(r^{n-2}f_n) = \frac{1}{nr}\frac{d\left(r^n f_{n+2}\right)}{dr}$$

Thus if we define the scaled wave function $F_n(r) = r^{n-2} f_n(r)$, we have

$$F_n = \frac{1}{nr}\frac{dF_{n+2}}{dr}$$

In other words, F_n is simply a scaled derivative of F_{n+2}, so this gives a sequence of solutions in *descending* space dimensions, by a formula very similar to equation.

At this point we should note that the negative sign in is potentially misleading, because the negation can be accomplished either by multiplying by negative 1 or by applying a phase shift to the periodic parts of the function, specifically by subtracting π from the phase. A generalized differentiation operator applied to the sine or cosine function amounts to a simple phase shift, and this just happens to give negation when the phase shift is π. Therefore, instead of writing equation with a leading factor of –1, we will write it with a leading factor of Φ^{-1}, where Φ is a linear phase shift operator defined by

$$\Phi\frac{g(\phi(x))}{x^k} = \frac{g(\phi(x)+\pi)}{x^k}$$

where g is a periodic function. Linearity means that $\Phi(a+b) = \Phi(a) + \Phi(b)$. For example,

$$\Phi\left(\frac{\sin(r)+r\cos(3r)}{r^2}\right)=\left(\frac{\sin(r+\pi)+\mathrm{r}\cos(3r+\pi)}{r^2}\right)$$

Using this operator, we will take as the general form of equation the expression

$$f_{n+2}=\Phi^{-1}\frac{n}{r}\frac{df_n}{dr}$$

We can make use of either to derive simple explicit expressions for elementary wave solutions in n dimensions, for both odd and even values of n. First, notice that in view of the identity we can express equations

$$2\frac{d}{d\left(x^2\right)}=\frac{1}{x}\frac{d}{dx}$$

and in the form

$$F_nc=\frac{2}{n}\frac{dF_{n+2}}{d\left(r^2\right)}f_{n+2}=2\Phi^{-1}n\frac{df_n}{r^2}$$

Given the elementary one-dimensional solution $F_1(r)$, the left hand relation allow us to determine a sequence of solutions by repeated integration as follows

$$F_3(\mathrm{r})=\frac{1}{2^{n+1}}\frac{(2n+1)}{2^n n!}$$

$$F_5(\mathrm{r})=\frac{3}{2}\int_{u^2=0}^{r^2}F_3(u)d\left(u^2\right)$$

$$F_5(\mathrm{r})=\frac{5}{2}\int_{u^2=0}^{r^2}F_5(u)d\left(u^2\right)$$

and so on. Now, in one-dimensional space we have the elementary solution $f_1(r) = cos(r)$, which in scaled form is $F_1(r) = cos(r)/r$. Therefore we have

$$F_3(r)=\frac{1}{2}\int_{u^2=0}^{r^2}F_1(u)d\left(u^2\right)=\frac{1}{2}\int_{u^2=0}^{r^2}\frac{\cos(u)}{\mathrm{u}}2u\,du=\sin(r)$$

and hence $f_3(r) = sin(r)/r$ as expected. Integrating again (with the appropriate scale factor) on F_3 gives F_5, and so on. Thus, using Cauchy's formula for repeated integration, we can express F_{2n+3} by performing the integration $n+1$ times. The product of the leading scale factors is $\left(\frac{1}{2}\right)\left(\frac{3}{2}\right)\left(\frac{5}{2}\right)\cdots\left(\frac{2n+1}{2}\right) = \frac{1}{2^{n+1}}\frac{(2n+1)}{2^n n!}$ so the repeated integrations give the formula

$$F_{2n+3}(\mathrm{r}) = \frac{(2n+1)}{2^{2n+1}(n!)^2}\int_{u^2=0}^{r^2}\left(r^2-u^2\right)^n F_1(u)\,d\left(u^2\right)$$

$$= \frac{(2n+1)}{2^{2n}(n!)^2}\int_{u=0}^{r}\left(r^2-u^2\right)^n \cos(u)\,du$$

Factoring an r^{2n} out of the integrand and making the substitution $x = u/r$, this can be re-written as

$$F_{2n+3}(r) = r^{2n+1}\,f_{2n+3}(r) = r^{2n+1}\frac{(2n+1)}{2^{2n}(n!)^2}\int_{x=0}^{1}\left(1-x^2\right)^n\cos(xr)\,dx$$

The natural generalization of this (n+1)th order integral is to allow n to be a non-integer n, and replacing the factorial of n with the gamma function of n+1. Thus we arrive at

$$F_{2\nu+3}(r) = \frac{\Gamma(2\nu+2)}{2^{2\nu}\Gamma(\nu+1)^2}\int_{x=0}^{1}\left(1-x^2\right)^{\nu}\cos(xr)dx$$

For space of two dimensions we set $\nu = -1/2$ in the above formula to give the solution

$$F_2(r) = \frac{\Gamma(0)}{2^{-1}\Gamma(1/2)^2}\int_{x=0}^{1}\left(1-x^2\right)^{-1/2}$$

$$\cos(xr)dx = \frac{2}{\pi}\int_{x=0}^{1}\frac{\cos(xr)}{\sqrt{1-x^2}}dx$$

This agrees with the result found in the note on Huygens' principle by means of a series evaluation, and we note that it is the lowest-order Bessel function, i.e., $f_2(r) = J_0(r)$.

Incidentally, with regard to equation, it's worth noting the identity

$$\int_{x=0}^{1}\left(1-x^{2}\right)^{\nu}dx = \frac{2^{2}\nu\Gamma\left(\nu+1\right)^{2}}{\Gamma\left(2\nu+1\right)}$$

Thus equation can be written in the form

$$f_{2n+3} = \frac{\int_{x=0}^{1}\left(1-x^{2}\right)^{\nu}\cos\left(xr\right)dx}{\int_{x=0}^{1}\left(1-x^{2}\right)^{\nu}dx}$$

which shows that the wave function is the mean of cos(xr) on the interval from $x = 0$ to 1 with the density proportional to $(1-x^2)^{\nu}$.

An alternative approach to developing wave solutions in n dimensions is to use the ascending sequence of solutions given by equation, beginning with the elementary one-dimensional solution $f_1(r) = cos(r)$, as follows

$$f_3(\mathrm{r}) = \Phi^{-1}\frac{1}{r}\frac{d}{dr}\left(\cos\left(r\right)\right) = \frac{\sin\left(r\right)}{r}$$

$$f_5(\mathrm{r}) = \Phi^{-1}\frac{3}{r}\frac{d}{dr}\left(\frac{\sin\left(r\right)}{r}\right) = 3\frac{\sin\left(r\right)-r\cos\left(r\right)}{r^{3}}$$

$$f_7(\mathrm{r}) = \Phi^{-1}\frac{5}{r}\frac{d}{dr}\left(3\frac{\sin\left(r\right)-\cos\left(r\right)}{r^{3}}\right)$$

$$= 15\frac{3\sin\left(r\right)-3r\cos\left(r\right)-r^{2}\sin\left(r\right)}{r^{5}}$$

and so on. Thus the ascent from n dimensions to $n+2$ dimensions is given by applying the differential operator

$$\Phi^{-1}\frac{n}{r}\frac{d}{dr} = 2\Phi^{-1}n\frac{d}{d\left(r^{2}\right)}$$

Applying this operator n times to a one-dimensional solution gives an expression for a $2n+1$ dimensional solution. So, on this basis, our expression for higher-dimensional solutions is

$$f_{2n+1}(\mathrm{r}) = \Phi^{-n}\frac{\left(2n!\right)}{\left(r^{2}\right)}\left[\frac{d}{d\left(r^{2}\right)}\right]^{n}\cos\left(r\right)$$

Since each differentiation (with respect to r^2) raises the number of space dimensions by two, we need to apply a half-order differentiation to produce solutions for spaces with even numbers of dimensions. Using the Left Hand Rule for fractional differentiation, we first perform a half-integration on *sin*(r) with respect to r^2 by means of Cauchy's formula, which gives

$$\frac{d^{-1/2}}{d\left(r^2\right)^{-1/2}}\cos(r) = \frac{1}{\Gamma(1/2)}\int_{u^2=0}^{r^2}\left(r^2-u^2\right)^{-1/2}\cos(u)\,d\left(u^2\right)$$

$$= \frac{2}{\sqrt{\pi}}\int_{u=0}^{r}\left(r^2-u^2\right)^{-1/2}\cos(u)\,u\,du$$

Dividing out the r^2, and making the substitution $x = u/r$, and noting that $du = r\,dx$, this can be written as

$$\frac{d^{-1/2}}{d\left(r^2\right)^{-1/2}}\cos(r) = \frac{2}{\sqrt{\pi}}\int_{u=0}^{1}\left(1-x^2\right)^{-1/2}\cos(xr)\,xr\,dx$$

We now need to perform one whole differentiation with respect to r^2, and we can do this inside the integration. Noting that

$$\frac{d}{d\left(r^2\right)}\left[xr\cos(xr)\right] = \frac{1}{2r}\frac{d}{dr}\left[xr\cos(xr)\right]$$

$$= x\frac{\cos(xr)-xr\sin(xr)}{2r}$$

we finally arrive at the half-derivative of *cos*(r)

$$\frac{d^{1/2}}{d\left(r^2\right)^{1/2}}\sin(r) = \frac{1}{r\sqrt{\pi}}\int_{u=0}^{1}\left(1-x^2\right)^{-1/2}$$

$$x\left[\cos(xr)-xr\sin(xr)\right]dx$$

Inserting this into equation with $n = 1/2$ gives

$$f_2(r) = \frac{2}{\pi r}\int_{u=0}^{1}\frac{\sin(xr)+xr\cos(xr)}{\sqrt{1-x^2}}x\,dx$$

This again equals the Bessel function $J_0(r)$, although the equivalence is non-trivial. Thus we can arrive at consistent results using the sequence of solutions related by differentiation in either the ascending or descending space dimensions.

The correspondences between solutions in higher and lower dimensions is essentially one-to-one, and all the solutions in higher dimensions can be mapped down to (or up from) a solution in one dimension. Using the series solution technique described in the note on Huygens' principle, it can be shown that all one-dimensional solutions are of the form

$$f_1(\mathrm{r}) = c_0 \cos(r) + c_1 \sin(r)$$

for some constants c_0 and c_1. So, returning to equation, generalizing to non-integer orders by replacing n with the real value ν and replacing the factorials with gamma functions, and inserting the general one-dimensional solution in place of the special solutions *cos*(r), we have

$$\mathrm{f}_{2n+1}(\mathrm{r}) = \Phi^{-\nu}\frac{\Gamma(2\nu+1)}{\Gamma(\nu+1)}\left[\frac{d}{d(r^2)}\right]^{\nu}\left[c_0\cos(r)+c_1\sin(r)\right]$$

The phase shift and differentiation operators are commutative, so we can apply the shift to the basic argument, and write this in the form

$$\mathrm{f}_{2n+1}(\mathrm{r}) = \frac{\Gamma(2\nu+1)}{\Gamma(\nu+1)}\left[\frac{d}{d(r^2)}\right]^{\nu}\left[c_0\cos(r-\nu\pi)+c_1\sin(r-\nu\pi)\right]$$

This shows that for any given number of space dimensions there is a two-parameter family of solutions. Also, for any choice of those two parameters we can vary the number of dimensions continuously.

Propagation of Pressure and Waves

Consider a linear sequence of N point-like particles, each of mass m, connected by springs with spring constants k as illustrated below for $N = 5$.

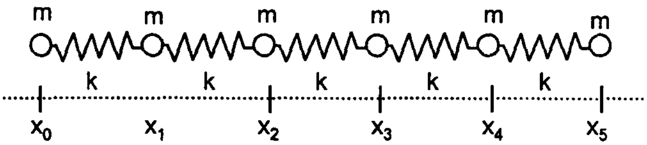


Fig. 14 Propagation of Waves

The position of the jth particle at any given time t is $x_j(t)$. The equilibrium length of each spring is L, and the particles are all initially

at rest at the locations $x_j(0) = jL$. We will take as a boundary condition that x_5 is rigidly held in its position, so $x_5(t) = 5L$ for all t. Also, the left-most mass will be driven so that $x_0(t)$ = vt beginning at the time $t = 0$. The equations of motion for the remaining four particles are

$$m\dot{x}_1 = k((x_2 - x_1) - L) - k\,((x_1 - x_0) - L)$$

$$m\ddot{x}_2 = k((x_3 - x_2) - L) - k\,((x_2 - x_1) - L)$$

$$m\ddot{x}_3 = k((x_4 - x_3) - L) - k\,((x_3 - x_2) - L)$$

$$m\ddot{x}_4 = k((x_5 - x_4) - L) - k\,((x_4 - x_3) - L)$$

Letting ω^2 denote the ratio k/m, and re-arranging terms, the equations can be expressed in matrix form as

$$\frac{1}{\omega^2}\begin{bmatrix}\ddot{x}_1\\ \ddot{x}_2\\ \ddot{x}_3\\ \ddot{x}_4\end{bmatrix} = \begin{bmatrix}-2 & 1 & 0 & 0\\ 1 & -2 & 1 & 0\\ 0 & 1 & -2 & 1\\ 0 & 0 & 1 & -2\end{bmatrix}\begin{bmatrix}x_1\\ x_2\\ x_3\\ x_4\end{bmatrix} + \begin{bmatrix}vt\\ 0\\ 0\\ 5L\end{bmatrix}$$

The parameter ω (equal to the square root of k/m) has units of rad/sec, and represents the characteristic rate of phase change for this system. It's easy to see that this system of linear differential equations has the particular solution $X(t) = P(t)$ where P is the column vector with components

$$P(t) = jL + \frac{5-j}{5}vt \quad \text{j=1, 2, 3 4}$$

so we just need to solve the homogeneous system to arrive at the complete solution. Letting M denote the coefficient matrix, the homogeneous system can be written symbolically as

$$\ddot{X} - w^2 MX = 0$$

so the eigenvalues can be expressed symbolically as $\pm\omega M^{1/2}$. However, determining the square root of a matrix is not trivial. A more practical approach is to solve for the squares of the eigenvalues of the original matrix equation. The trial solution $x_j(t) = A_j e^{\lambda t}$, where A_j has units of length and l has units of time^{-1}, leads to the characteristic equation

$$det\begin{bmatrix} -2\left(\frac{\lambda}{\omega}\right)^2 & 1 & 0 & 0 \\ 1 & -2-\left(\frac{\lambda}{\omega}\right)^2 & 1 & 0 \\ 0 & 1 & -2-\left(\frac{\lambda}{\omega}\right)^2 & 1 \\ 0 & 0 & 1 & -2-\left(\frac{\lambda}{\omega}\right)^2 \end{bmatrix}$$

$$=\left(\left(\frac{\lambda}{\omega}\right)^2\right)^4+8\left(\left(\frac{\lambda}{\omega}\right)^2\right)^3+21\left(\left(\frac{\lambda}{\omega}\right)^2\right)^2+20\left(\left(\frac{\lambda}{\omega}\right)^2\right)+5$$

The quartic in $(\lambda/\omega)^2$ factors into two quadratics, which can be solved to give the eight purely imaginary eigenvalues of the system

$$\lambda_{1,5}=\pm i\omega\sqrt{\frac{3-\sqrt{5}}{2}} \qquad \lambda_{2,6}=\pm i\omega\sqrt{\frac{3+\sqrt{5}}{2}}$$

$$\lambda_{3,7}=\pm i\omega\sqrt{\frac{5-\sqrt{5}}{2}} \qquad \lambda_{4,8}=\pm i\omega\sqrt{\frac{5-\sqrt{5}}{2}}$$

Letting ρ_j denote $\lambda_j/(i\omega)$, we can take advantage of the fact that the eigenvalues come in conjugate imaginary pairs to express the solution of the homogeneous equation in the form

$$x_j(t)=\sum_{k=1}^{4}\left(c_{j,k}e^{i\omega\rho_1 t}+c_{j,k+4}e^{-i\omega\rho_1 t}\right)$$

where the coefficients $c_{j,k}$ are complex constants. This can be expressed as a sum of real-valued sine and cosine functions of real arguments, but we find that the coefficients of the cosine terms must all vanish (to satisfy the conditions of the problem), so we are left with an expression of the form

$$x_j(t)=\sum_{k=1}^{4}A_{j,k}\sin\left(\omega\rho_k t\right)$$

So, combining the homogeneous solution with the particular solution, we know that each of the mass particle positions is of the form $x_j(t)=A_{j,1}sin(r_1wt)+A_{j,2}sin(r_2wt)+A_{j,3}sin(r_3wt)$

$$+ A_{j,3}\sin(r_4 wt) + jL + \frac{N-j}{N}vt$$

where the coefficients $A_{j,k}$ are to be determined by the initial conditions. (The cosine terms are all zero.) Making use of the sequence of differentiated system equations

$$\ddot{X} = \omega^2 MX + U\left\{+\frac{N-j}{N}v \; for\, n=1\right\} = \omega^2 M\dot{X} + U\ddddot{X} = \omega^2 M\ddot{X} \; etc.$$

and letting $x_j^{(n)}$ denote the nth derivative of $x_j(t)$ at t = 0, we can construct the following table of initial conditions:

n	$x_0^{(n)}$	$x_1^{(n)}$	$x_2^{(n)}$	$x_3^{(n)}$	$x_4^{(n)}$	$x_5^{(n)}$
0	0	1	2	3	4	5
1	v	0	0	0	0	0
2	0	0	0	0	0	0
3	0	$\omega^2 v$	0	0	0	0
4	0	0	0	0	0	0
5	0	$-2\omega^4 v$	$\omega^4 v$	0	0	0
6	0	0	0	0	0	0
7	0	$5\omega^6 v$	$-4\omega^6 v$	$\omega^6 v$	0	0
8	0	0	0	0	0	0
9	0	$-14\omega^8 v$	$14\omega^8 v$	$-6\omega^8 v$	$\omega^8 v$	0
10	0	0	0	0	0	0

Our general solution automatically satisfies the even-ordered derivatives, because we have set all the coefficients of the cosine terms to zero, so we need only equate the odd-ordered derivatives in the table to the corresponding derivatives of the general solution at t = 0

$$x_j^{(n)}(0) = A_{j,1}(r_1 w)^n + A_{j,2}(r_2 w)^n + A_{j,3}(r_3 w)^n + A_{j,4}(r_4 w)^n$$

$$\left\{+\frac{N-j}{N}v \; for\, n=1\right\}$$

This leads to the system of equations

$$\begin{bmatrix} \omega\rho_1 & \omega\lambda_2 & \omega\lambda_3 & \omega\lambda_4 \\ (\omega\rho_1)^3 & (\omega\lambda_2)^3 & (\omega\lambda_3)^3 & (\omega\lambda_4)^3 \\ (\omega\rho_1)^5 & (\omega\lambda_2)^5 & (\omega\lambda_3)^5 & (\omega\lambda_4)^5 \\ (\omega\rho_1)^7 & (\omega\lambda_2)^7 & (\omega\lambda_3)^7 & (\omega\lambda_4)^7 \end{bmatrix}$$

$$[A]^T = \begin{bmatrix} \frac{-4v}{5} & \frac{-3v}{5} & \frac{-2v}{5} & \frac{-v}{5} \\ -\omega^2 v & 0 & 0 & 0 \\ -2\omega^2 v & \omega^4 v & 0 & 0 \\ -5\omega^2 v & 4\omega^6 v & -\omega^6 v & 0 \end{bmatrix}$$

The *j*th row of the left-hand matrix is a multiple of w^{2j-1}, whereas the *j*th row of the right-hand matrix is a multiple of w^{2j-2}. Also, the right hand matrix is a multiple of *v*. Therefore, if we solve this system by bringing the inverse of the leading matrix over to the right side, we can eliminate all the appearances of *v* and *w* in the matrices, and simply apply a factor of *v*/*w* to the result, so the transpose of the coefficient matrix *A* is given by

$$\begin{bmatrix} A_{11} & A_{21} & A_{31} & A_{41} \\ A_{12} & A_{22} & A_{32} & A_{42} \\ A_{13} & A_{23} & A_{33} & A_{43} \\ A_{14} & A_{24} & A_{34} & A_{44} \end{bmatrix} = \frac{v}{\omega} \begin{bmatrix} \rho_1 & \rho_2 & \rho_3 & \rho_4 \\ \rho_1^3 & \rho_2^3 & \rho_3^3 & \rho_4^3 \\ \rho_1^5 & \rho_2^5 & \rho_3^5 & \rho_4^5 \\ \rho_1^7 & \rho_2^7 & \rho_3^7 & \rho_4^7 \end{bmatrix}^{-1}$$

$$\begin{bmatrix} -4/5 & -3/5 & -2/5 & -1/5 \\ -1 & 0 & 0 & 0 \\ -2 & 0 & 0 & 0 \\ -5 & 4 & -1 & 0 \end{bmatrix}$$

Recall that the values of ρ_j are solutions of

$$(\rho^2)^4 - 8(\rho^2)^3 + 21(\rho^2)^2 - 20(\rho^2) + 5 = 0$$

As discussed in Linear Fractional Transformations, polynomials of this type (with coefficients from diagonals of Pascal's triangle) have the trigonometric solution $r_j = 2cos\left(\frac{j\pi}{2N}\right)$, $j = 1, 2, ..., N-1$ where $N = 5$ in our example. (We take just one of the square roots of each root ρ^2 of the (*N*–*1*)th degree polynomial.) Thus if we let *R* denote the matrix whose inverse is taken in the above equation, we can state the components of *R* explicitly as

$$R_{i,j} = \left(2cos\left(\frac{j\pi}{2N}\right)\right)^{2i-1}$$

Obviously the components of R are dimensionless. The right-most factor in the preceding matrix equation, which we will denote by C, represents the definition of the initial conditions and the particular solution. The elements of the first row are simply $(N\text{-}j)/N$, and the magnitudes of the elements on the remaining rows can be generated recursively by the relation $C_{m+1,\,n} = -C_{m,\,n-1} + 2C_{m,\,n} - C_{m,\,n+1}$

Thus after the first row we append the first N–2 rows and N–1 columns of the array

–1	0	0	0	0	0	0	0
–2	1	0	0	0	0	0	0
–5	4	–1	0	0	0	0	0
–14	14	–6	1	0	0	0	0
–42	48	–27	8	–1	0	0	0
–132	165	–110	44	–10	1	0	0
–429	572	–429	208	65	12	–1	0
–1430	2002	–1638	910	–350	90	–14	1

etc

The numbers in the first column are obviously the (negative) Catalan numbers, as are the sums of the numbers in each row. Notice that the recurrence relation applies to the first row with the fractional terms as well, and can be exercised in reverse to generate the infinite sequence of previous rows. The components of C are, of course, dimensionless.

In terms of the matrices defined above the positions of the N–1 mass particles as a function of time are given by the row vector

$$X^T(t) = P^T(t) + \frac{v}{\omega}E(t)R^{-1}C$$

where $E(t)$ is the dimensionless row vector with the components

$$E_j(t) = \sin\left(2\cos\left(\frac{j\pi}{2N}\right)\omega t\right) \qquad j = 1, 2, \ldots, N-1$$

Recall that $P_j(t) = jL + (1 - j/N)$vt, so if we express the speed v in the form nL where n is the number of 'L-distances' per unit time, and if we note that the homogeneous part of the solution is also a multiple of v, we can divide through by L to give the fully dimensionless equation

$$\frac{X^T(t)}{L} = \frac{P^T(t)}{L} + \frac{v}{\omega}E(t)R^{-1}C$$

where the elements of the first term on the right side are $j + (1\text{-}j/N)nt$. The coefficient (n/ω) is the ratio of two parameters, each with units of time^{-1}, but the parameter n signifies a number of L-distances moved by x_0 per unit time, whereas ω represents a number of phase radians per unit time. Multiplying either of these by the length L gives something with units of speed. The quantity Ln equals the speed v of x_0, and the quantity $L\omega$ is a characteristic speed related to the phase of the system itself. We will denote this speed by c, and refer to it as the acoustic speed of the system, because it is the speed at which pressure disturbances propagate through the system. To see this, recall that the speed of sound in a material medium is

$$C = \sqrt{\frac{dp}{d\rho}}$$

where p is the pressure and ρ is the density. Our mass-spring system is just one-dimensional, but we can arbitrarily assign it a cross-sectional area of A, and we can let x denote a small deviation in the length of a spring from its null-force length L. In these terms, the pressure (force per area) is $p = kx/A$ and the density (mass per volume) is $\rho = m/(A(L–x))$. From this we have

$$\frac{dp}{dx} = \frac{k}{A} \quad \frac{d\rho}{dx} = \frac{m}{A(L-x)^2} \approx \frac{m}{AL^2}$$

and therefore

$$c = \sqrt{\frac{dp}{d\rho}} = L\sqrt{\frac{k}{m}} = Lw$$

To give an intuitive idea of why $L\omega$ should be the phase speed for wave propagation, consider an infinite sequence of mass-springs, and suppose each mass particle is in steady sinusoidal motion, oscillating about its null position, so the particles are always separated by a distance close to the null distance L. The equation of motion for each particle is of the form

$$m\ddot{x}_j = k(x_{j-1} - 2x + x_{j+1})$$

Letting ϕ denote the uniform phase shift from one mass to the next, i.e., the phase shift over a distance L, we can put

$$x_{j-1}(t) = sin(Wt - f)$$
$$x_j(t) = sin(Wt)$$
$$x_{j+1}(t) = \sin(Wt + f)$$

Substituting into the equation of motion and simplifying, we get

$$\Omega^2 = 2(1 - \cos(f))\frac{k}{m} = 4\sin(\phi/2)^2\frac{k}{m}$$

and therefore, since ϕ is small because two adjacent particles will not be far out of phase, we have

$$\Omega = 2\sin(\phi/2)\sqrt{\frac{k}{m}} \approx \phi\sqrt{\frac{k}{m}} = \phi\omega$$

Dividing through by ϕ gives Ω/ϕ, which represents the phase change, expressed in units of time^{-1}, over the distance L from one particle to the next. Therefore, $(\Omega/\phi)L = \omega L = c$ is the phase velocity, in agreement with the fluid-mechanical derivation.

As one would expect, if the speed v of particle x_0 is small compared with c, the result is a quasi-static compression of all the particles, for the case $N = 5$ with $v/c = 1/50$ and $v/c = 1/20.$

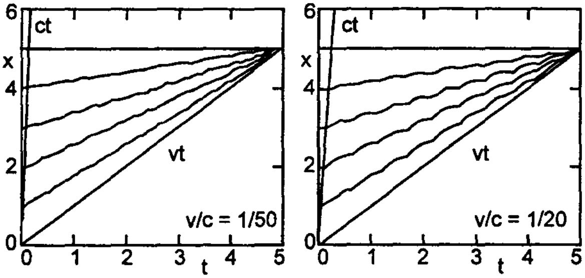


Fig. 15

However, as the ratio of v/c increases, we can begin to see the dynamic propagation delay. The figure are for $v/c = 1/10$ and $v/c = 1/5.$

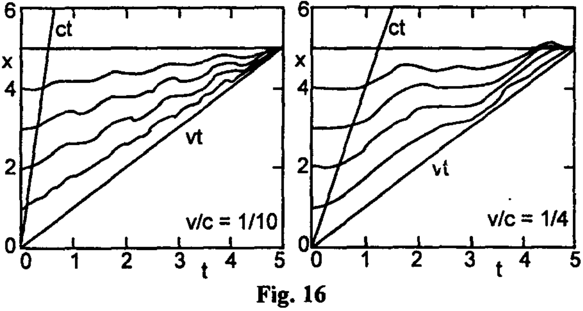


Fig. 16

In each case the particles are not appreciably affected until nearly the time when the '*ct*' line reaches them. In other words, the compression effect propagates at the same speed c as do acoustic waves.

Incidentally, it's possible for the particles to pass each other dynamically, because the restorative spring force is null for a mutual distance of one unit length, and varies linearly away from that condition. There is nothing singular about the zero-length condition of these idealized springs. A different type of model could be based on, say, mutual inverse-square repulsion, which goes to infinity as the separation goes to zero, so the particles could never pass each other. However, classical force laws of that type do not apply just to neighbouring particles but to all the other particles, as instantaneous forces at a distance, so for purposes of illustrating how pressure propagates strictly by 'contact forces' it is more convenient to represent the mutual forces as springs (not to mention the fact that the infinite potentials of ideal point-like particles are probably not realistic either).

As we increase the ratio of v/c still further, we continue to see that the pressure propagates at essentially the speed c, the cases $v/c = 1/2$ and $v/c = 1$.

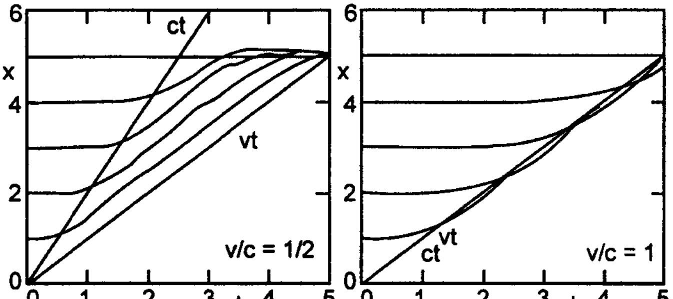


Fig. 17

The fact that there is almost no response at the jth mass particle until $x_j(0)/c$ after the particle at $x = 0$ begins to move might seem counter-intuitive at first, because we know the jth particle begins to move as soon as the $(j-1)$th particle begins to move, so they should all begin to move at time $t = 0$. Of course, this supposition relies on our assumption that each spring transmits force instantaneously as a function of the distance between its endpoints. To model a realistic spring we would need to account for the finite acoustic propagation speed of the spring itself, treating each small part of the spring as an element with a certain mass and restorative force. But we have not done this, so there ought to be some instantaneous action at a distance, and indeed if we examine the initial time period closely, by plotting $ln(x_j(t) - j)$ versus t and $ln(t)$, we can see that the motion does begin for all the particles at $t = 0$. (The plots below are for $v/c = 1/4$.)

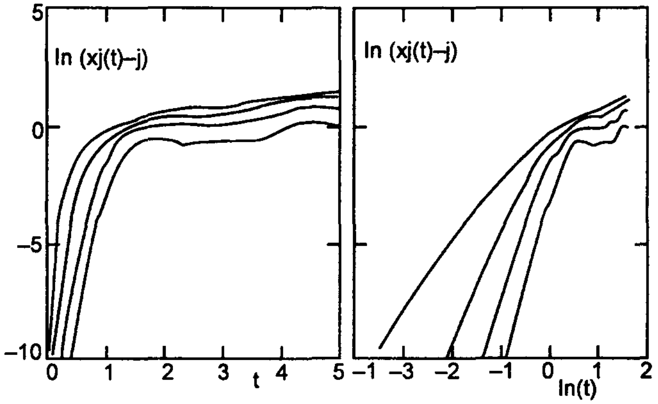


Fig. 18

These plots also show that the motion of the *j*th particle is exponentially small until a characteristic time that is proportional to the distance from the source of the disturbance, consistent with the fact that the jth particle is virtually unmoved until $x_j(0)/c$ after the disturbance begins. This is true even though we are modeling each spring as an instantaneous force transmitter.

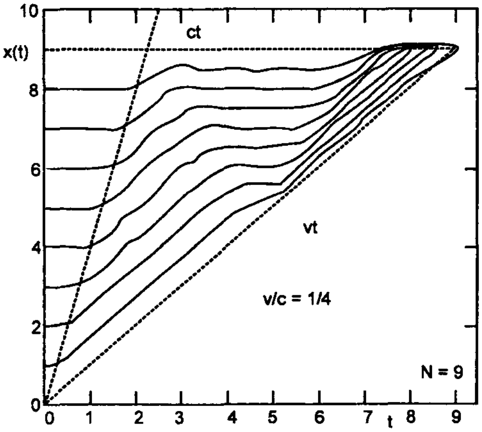


Fig. 19

If we increase the number of particles and springs we will gradually approach a truly continuous medium in which the distances over which the forces propagate instantaneously approach zero. For example, the case $N = 9$.

Again we see that the disturbance essentially propagates at the acoustic speed c, but now cutoff at the ct line is even sharper. We can show this by comparing the response of the 6th particle in the $N = 9$ case with the response of the 2nd particle in the $N = 3$ case, normalized to the same total distance.

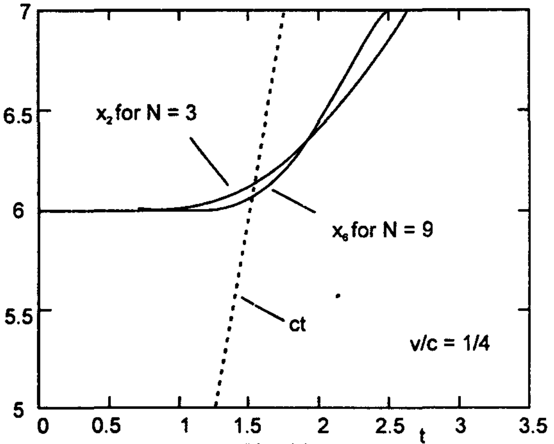


Fig. 20

As expected, the position $x_6(t)$ with $N=9$ makes a much sharper corner at the acoustic propagation line than does the position $x_3(t)$ with $N = 3$. As we continue to increase N (the number of spring-mass elements into which we divide the overall distance), the corner becomes progressively sharper. In the limit as N goes to infinity–which represents the situation in which all instantaneous force-at-a-distance has been eliminated–the response approaches perfect flatness until reaching the acoustic propagation line.

Thus, despite the fact that there is some non-zero instantaneous response (albeit fantastically small) in the discrete model for any finite N, no matter how large, the acoustic propagation speed becomes an absolute limit as those segments are reduced to zero. One implication is that, if we take the speed of light as an absolute limit on the propagation speed for any energy or information, then the speed limit must apply down to infintessimal scales. On the other hand, the non-zero probability amplitude for a photon to traverse a very small distance at a speed greater than *c* (according to quantum electrodynamics) could be interpreted as a propensity for 'action at a distance' over those small distances.

Maxwell included some interesting comments on this subject in his *Treatise on Electricity and Magnetism.* After deriving the general time-dependent equations for electromagnetic disturbances in terms of the parameters *C, K*; and μ (the specific conductivity, the specific capacity for electrostatic induction, and the magnetic permeability of the medium, respectively), he considers the propagation of such disturbances in two different limiting cases. First, he considers propagation with $C = 0$, i.e., in a non-conducting medium (of which the vacuum would be one example), and shows that the disturbances propagate at the speed

$$V = \frac{1}{\sqrt{K\mu}}$$

It so happens that the numerical value of this expression equals the speed of light. It was the crowning achievement of Maxwell's electrodynamic theory that he was able to derive the speed of light in terms of these parameters of electricity and magnetism. Then in Article 801 he considers "the case of a medium in which the conductivity is large in proportion to the inductive capacity. In this case we may leave out the term involving *K* in the equations of Article 783, and they then become

$$\nabla^2 F + 4\pi\mu C\frac{dF}{dt} = 0$$

[and the same for the other components]. Each of these equations is of the same form as the equation of the diffusion of heat given in Fourier's *Traite de la Chaleur.*" Maxwell then goes on to discuss the analogy between heat transfer and the diffusion of electromagnetic quantities. For an infinite medium whose initial conditions are known, Fourier had already solved this equation. The value of *F* at any given point at the time *t* is the weighted average of the values at every other point, where the weight assigned to a point at a distance *r* is

$$e^{-\pi\mu}Cr^2 / t$$

Thus at the initial time $t = 0$ each point just has its own arbitrarily defined value, because the weights for all other point with *r* greater than 0 are zero. As time increases, the radius r for which there is a significant weight increases–in proportion to the square root of the time. The exponential dependence on time corresponds to the linearity of the logarithmic plots shown above for the motion of the mass particles ahead of the acoustic speed. Then Maxwell makes the interesting remarks.

> There is no determinate velocity which can be defined as the velocity of diffusion. If we attempt to measure this velocity by ascertaining the time requisite for the production of a given amount of disturbance at a given distance from the origin of disturbance, we find that the smaller the selected value of the disturbance the greater the velocity will appear to be, for however great the distance, and however small the time, the value of the disturbance will differ mathematically from zero. This peculiarity of diffusion distinguishes it from wave-propagation, which takes place with a definite velocity. No disturbance takes place at a given point till the wave reaches that point, and when the wave has passed, the disturbance ceases for ever.

This is reminiscent of how, with our mass-spring 'diffusion' system, if we examine the initial portion of $x_j(t)$ more and more closely to determine precisely when it begins to change, we find that it has non-zero change for all t greater than 0, although the magnitude of the change is exponentially small prior to the delay time of D/c. Thus, just as Maxwell says, if we define our threshold small enough, the speed of propagation can be as great as we choose. However, this is only because our model contains implicit action-at-a-distance elements. As noted above, each spring is considered to exert equal and opposite forces at both ends strictly as a function of the difference between the instantaneous positions of the ends. In the limit of a pure contact medium with no extended instantaneous elements, this effect disappears, and the acoustic speed limit becomes absolute for the propagation of any disturbance. It's odd that Maxwell should have regarded the lack of an absolute speed limit in his artificial 'diffusion' example as having physical significance, because he had already shown that the propagation speed for waves was inversely proportional to the square root of K, whereas in the diffusion example he explicity applies the approximation $K = 0$, i.e., he *assumes* in this case that the speed of light is infinite, so it should come as no surprise that there is no upper bound on the speed of diffusion under this assumption. Had he worked out the speed of diffusion for non-zero *K*, he would have found that it does. not exceed the speed of wave propagation, as illustrated by the simple mass-spring model.

This highlights the profound qualitative difference between the propagation of disturbances in a model with an arbitrarily large (yet finite) number of "distant-action" springs versus the propagation in a continuous model. The former has (in principle) no upper bound on propagation speed, whereas the latter exhibits a strict speed limitation of *c* for the propagation of any disturbance. The reason for this profound difference involves subtle aspects of mathematical limits, convergence, and existence of functions–the very issues that Fourier is often accused of having overlooked in his treatment of heat flow by means of Fourier series. We can explain these issues by examining the power series expressions for $x_j(t)$. Recall our table of derivatives of these functions at $t = 0$. If we consider $x_j(t)$ as a power series in *t* with constant coefficient, then the *n*th derivative is *n*! times the coefficient of t^n. Thus we have the following power series for the normalized position functions:

$$\mathrm{x}_0(\mathrm{t}) - 0 = \frac{v}{c}(\omega t)$$

$$\frac{x_1(t)}{L} - 1 = \left(\frac{1}{3!}(\omega t)^3 - \frac{2}{5!}(\omega t)^5 + \frac{5}{7!}(\omega t)^7 \right.$$
$$\left. - \frac{14}{9!}(\omega t)^9 + \frac{42}{11!}(\omega t)^{11} - \ldots\right)$$

$$\frac{x_2(t)}{L} - 2 = \frac{v}{c}\left(\frac{1}{5!}(\omega t)^5 - \frac{4}{7!}(\omega t)^7 + \frac{14}{9!}(\omega t)^9 - \frac{48}{11!}(\omega t)^{11} - \ldots\right)$$

$$\frac{x_3(t)}{L} - 3 = \frac{v}{c}\left(\frac{1}{7!}(\omega t)^7 - \frac{6}{9!}(\omega t)^9 + \frac{27}{11!}(\omega t)^{11} - \ldots\right)$$

The lowest-degree term of each successive function is two powers of *wt* above that of the previous function, which corresponds to the fact that each successive function differs significantly from zero only at progressively larger values of *wt*. The 'knee' of each curve occurs when *wt* exceeds *j*. Multiplying both quantities by *L*, replacing ωL with *c*, and dividing both quantities by *c*, we find that the normalized position function differs significantly from zero only when *t* exceeds *jL/c*.

Nevertheless, each of these functions is, strictly speaking, nonzero for all positive values of *t*. In other words, for any finite *j* there is instantaneous action. But what about the limit as *N* goes to infinity, and not just countable infinity but a continuum? In that case the normalized position function for a particle at the distance *D* from the origin has no finite-degree term, i.e., it is rigorously zero until the time *D/c*, so there is no instantaneous action at a distance. This is an example of the 'limit paradox', which is resolved by noting that the limit of a set need not be an element of the set, and need not share all properties of the elements of that set.

Forces and Waves

Consider an illuminated charged sphere resting at the origin of a system *x,y,z* of inertial coordinates, and a small test particle moving with speed *v* in the positive *x* direction at a fixed *y* coordinate.

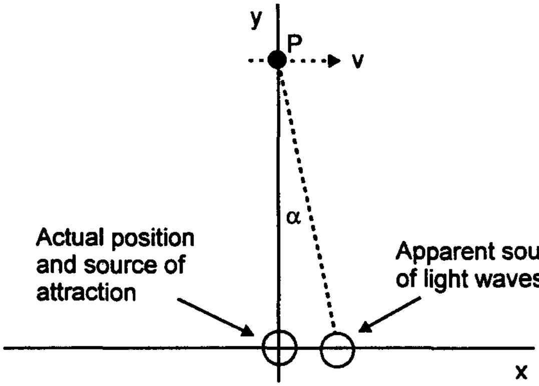

y
P
v
α
Actual position and source of attraction
Apparent source of light waves
x

Fig. 21 Aberration of Forces

If the distance between the two objects is sufficiently great, the light (electromagnetic waves) emanating from the sphere will consist of essentially planar horizontal waves when it reaches the test particle. Since the particle is moving tangentially with speed *v*, the angle of the incoming light will be affected by aberration, such that the apparent source of the light (from the point of view of the test particle as it crosses the *y* axis) is at an angle $\alpha = \arcsin(v/c)$ ahead of the actual position of the sphere. However, the direction of the electrical force exerted by the sphere on the test particle points directly toward the actual position of the sphere. Thus, the incoming electromagnetic waves from the sphere experience aberration, but the electromagnetic force of attraction to the sphere does not. This sometimes misleads people into thinking that the force somehow propagates instantaneously (to account for the absence of aberration).

The test particle to be at rest, and the charged sphere to be moving in the negative x direction with speed v. From this point of view, if D denotes the distance from the sphere to the particle, then at any time t the particle 'sees' the sphere at the location it occupied at a time $t - D/c$, because D/c is how long it take for light to travel the distance D.

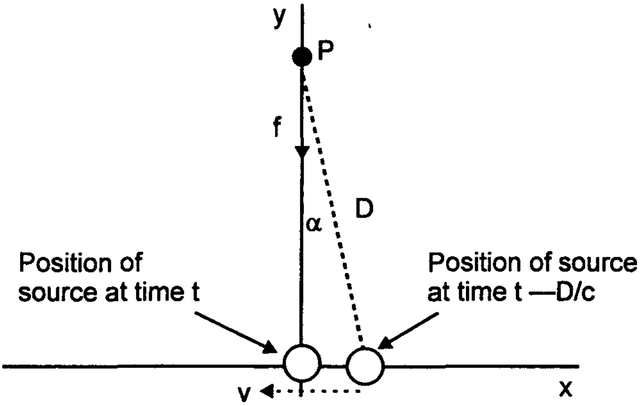

y
P
f
α
D
Position of source at time t
Position of source at time t —D/c
v
x

Fig. 22 Aberration of Waves.

Just as before, the light arrives at the (stationary) particle P from a direction differing from the true current direction of the source at time t by the angle α. Also, since we have simply changed coordinate systems, which can have no effect on any physical attributes, we know the electromagnetic force on the particle at the time t points directly toward the sphere's actual (not apparent) position at that time.

The absence of aberration in the direction of the electromagnetic force does not indicate that the force propagates infinitely fast. (In fact, the concept of a 'moving force' is not even well defined.) The force on a test particle at any given instant is due to the electromagnetic field in the immediate vicinity of the particle at that instant. In general the field at any given place and time consists of contributions from multiple sources at a variety of distances. The number of sources and their distances matter only insofar as they determine the electromagnetic field. The field of the charged sphere with respect to the rest frame of the sphere is an electro-static configuration (no magnetic field) with spherical symmetry centred on the source. A uniformly moving charged test particle in this field is subjected to a force proportional to (and therefore pointing in the same direction as) the electric field vector at its present location, so the force obviously points directly towards the source at all times.

On the other hand, in terms of the rest frame of the test particle the charged sphere is in uniform motion and the electromagnetic field has both electric and magnetic components. However, since the test particle is at rest with respect to these coordinates, it does not experience any magnetic force, so again the force on the particle is proportional to the electric field vector. To determine the direction of this force we need to know how the components of the electric field transform from one system of inertial coordinates to another. As explained in Force Laws and Maxwell's Equations, if E_x, E_y, and E_z are the components of the electric field at a given point with respect to the *x,y,z,t* coordinates, then the components with respect to a similarly oriented system of inertial coordinates *x',y',z',t'* moving with speed *v* in the positive *x* direction are

$$E_{x'} = E_x E_{y'} = \frac{E_{y'} - vB_z}{\sqrt{1-v^2}} E_{z'} = \frac{E_z - vB_y}{\sqrt{1-v^2}}$$

Of course, we also have

$$x' = \frac{x - vt}{\sqrt{1-v^2}} y' = y z' = z$$

In the unprimed coordinates (the rest frame of the charged sphere) we know the electric field components at the location of the test particle point directly toward the origin, which means E_x, E_y, and E_z are proportional to the coordinates *x*, *y*, and *z* of the test particle. Also, since the magnetic field is zero with respect to the unprimed coordinates, and since the origins of the two coordinate systems coincide at $t = 0$, we have

$$\frac{E_{x'}}{E_{y'}} = \frac{E_x}{E_y}\sqrt{1-v^2} = \frac{x}{y}\sqrt{1-v^2} = \frac{x'}{y'}$$

Similarly it follows that $E_{x'}/E_{z'} = x'/z'$ and $E_{y'}/E_{z'} = y'/z'$, confirming that the electric field vector at every location points directly toward the instantaneous source with respect to the rest frame of the test particle (relative to which the source is moving with a speed v). Thus the absence of 'force aberration' for objects in fully developed inertial motion is an immediate consequence of Lorentz covariance.

The qualifier 'fully developed' is necessary, because every object is instantaneously at rest with respect to some inertial frame, but the object's field in its current rest frame is spherical and satisfies the steady-state relations only out to a distance $D = c\Delta t$ where Dt is the length of time the object has been unaccelerated. This highlights the fact that although the field of an object exists and acts at a distance from the object, changes in the field propagate at the finite speed c. When an object changes its state of motion, the field must change accordingly, and these changes propagate outward from the source at the speed c. In the far field these changes propagate in the form of waves.

One thing that sometimes puzzles people about the lack of force aberration is that they tend to regard the electric field as the gradient of a potential, and they know the equi-potential surfaces for a uniformly moving charged particle are contracted in the direction of motion so they form ellipsoids instead of spheres, and clearly the spatial gradient of this potential does not point towards the centre (except for lines parallel or perpendicular to the axis of motion). The explanation is that the electric field vector equals the spatial gradient of the potential field only if the field is stationary, i.e., unchanging with time. If the field is changing with time, the full expression for the electric field must include an additional term to account for this, i.e., we have

$$E = -\nabla f - \frac{1}{c}\frac{\partial A}{\partial t}$$

Where A signifies the vector potential of the electromagnetic field. The second term on the right hand side represents the effect of the changing potential with time. Using the Lorentz gage

$$\nabla \cdot A = -\frac{1}{c}\frac{\partial \phi}{\partial t}$$

the field equations for the electromagnetic potentials are

$$\nabla^2\phi - \frac{1}{c^2}\frac{\partial^2\phi}{\partial t^2} = -4\mathrm{pr}$$

$$\nabla^2 A - \frac{1}{c^2}\frac{\partial^2 A}{\partial t^2} = -4\mathrm{pr}\,\frac{v}{c}$$

It follows that, if v is constant (and has been for a sufficiently long time), and we are given a solution $\phi(x,y,z,t)$ for the scalar potential, we can multiply this solution by v/c to give a solution of the vector potential

$$A = \frac{\phi}{c}v$$

Therefore, under these conditions, the time-dependent term in the last equation for E can be written as

$$\frac{1}{c}\frac{\partial A}{\partial t} = \frac{v}{c^2}\frac{\partial\phi}{\partial t}$$

Now, by definition, the total derivative of ϕ along any incremental path dx,dy,dz,dt is

$$df = \frac{\partial\phi}{\partial x}dx + \frac{\partial\phi}{\partial y}dy + \frac{\partial\phi}{\partial z}dz + \frac{\partial\phi}{\partial t}dt$$

Dividing by dt and solving for the partial of ϕ with respect to t gives

$$\frac{\partial\phi}{\partial t} = \frac{\partial\phi}{\partial t} - \left(\frac{\partial\phi}{\partial x}\frac{\partial x}{\partial t} + \frac{\partial\phi}{\partial y}\frac{\partial y}{\partial t} + \frac{\partial\phi}{\partial z}\frac{\partial z}{\partial t}\right)$$

Taking dx/dt etc., as the components of the sphere's velocity v, the total derivative $d\phi/dt$ represents the change in f with time along a co-moving worldline, and since the (fully developed) field is stationary with respect to the rest frame of the sphere, we have $d\phi/dt = 0$. Therefore the partial of f with respect to t equals the negative of the dot product of the spatial gradient of ϕ with the velocity v, so the previous expression for the time-dependent electric field is

$$E = -\nabla\phi + \frac{v}{c^2}(v\cdot\nabla\phi)$$

We are considering the case when the sphere's motion is in the positive x direction, so we have $\boldsymbol{v}=(v,0,0)$ and the above expression becomes.

$$E=-\left[\left(1-\frac{v^2}{c^2}\right)\frac{\partial\phi}{\partial x},\frac{\partial\phi}{\partial y},\frac{\partial\phi}{\partial z}\right]$$

A surface of constant ϕ is a stationary sphere in the rest frame of the source, so it transforms to an ellipsoid due to contraction in the direction of motion.

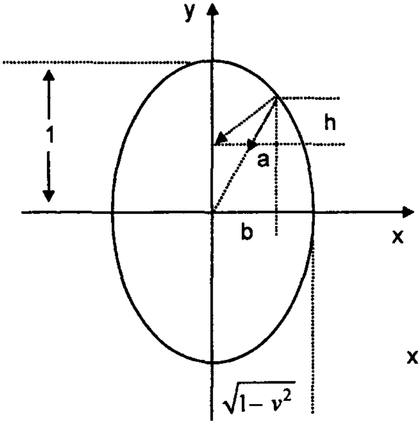


Fig. 23

The equation of this cross-section is

$$\left(\frac{x}{\sqrt{1-v^2}}\right)^2+y^2=1$$

where we have chosen units so that $c = 1$. Taking the differential of both sides gives

$$\frac{2x}{1-v^2}dx+2y\,dy=0$$

The slope of the normal to the ellipse at the point (x, y) is the negative reciprocal of dy/dx, which is

$$-\frac{dx}{dy}=\frac{h}{b}=(1-v^2)\frac{y}{x}$$

According to our expression for $E(r,t)$, we begin with the gradient of f and then reduce the x component by the factor $(1 - v^2)$, where we still have $c = 1$. Thus we have

$$\frac{h}{a}=(1-\mathrm{v}^2)\frac{y}{x\left(1-v^2\right)}=\frac{y}{x}$$

This confirms that the electromagnetic force exerted by the field of the moving charged sphere on the test particle at time *t* is directed toward the position of the sphere at the same time *t*. This is a natural consequence of Lorentz covariance, and does not imply any instantaneous transfer of energy or information.

It's true that, in quantum theory, the electromagnetic force can be considered to be mediated by photons, but these are *virtual* photons, which are actually just analytical components of the field. In effect these virtual photons form a cloud around the source particle, and they 'exist' only within the uncertainty envelope. An electromagnetic interaction between two electrons, for example, is modeled as an exchange of photons between the overlapping fields of two particles. It is not represented by a photon traversing from one particle to the other. Virtual photons don't even possess definite trajectories through space and time. They are conceptual entities arising in the quantization of the electromagnetic field.

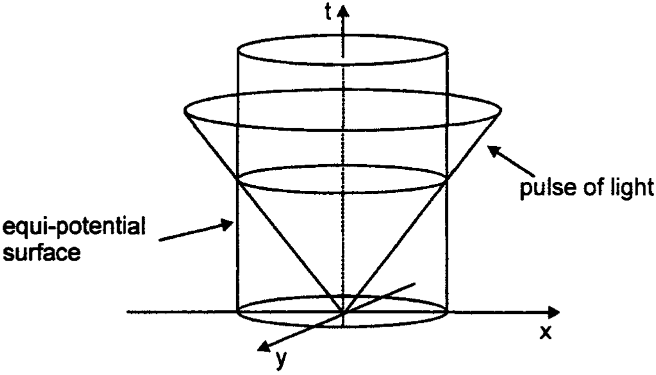


Fig. 24 Equipotential surface

Another point that sometimes puzzles people is why an equi-potential sphere transforms to an ellipsoid under a Lorentz transformation, whereas a spherical wave of light transforms to a spherical wave under the same transformation. The reason is that an equi-potential sphere is stationary, whereas a wave of light is expanding.

A side view showing the intersections of these two surfaces with two difference planes of simultaneity.

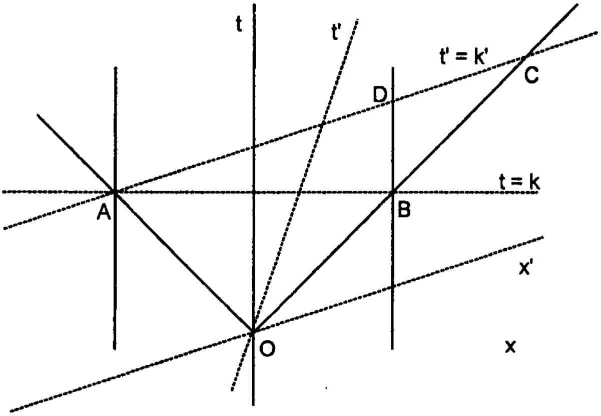


Fig. 24 Two Diffrerent Planes.

The source of the light pulse and the potential field moves along the *t* axis. With respect to the *x*, *t* coordinate system the expanding spherical shell of light coincides with an equi-potential sphere at the time $t = k$ with the diameter *AB*. However, with respect to the *x'*, *t'* coordinate system the left-most point of the expanding sphere of light just touches the left-most point of the equi-potential sphere at the point *A* and time $t' = k'$. At this time the right-most point of the light sphere is at *C*, whereas the right-most point of the equi-potential surface is at *D*. The 'centre' of the expanding light sphere (with respect to the primed coordinates) moves along the *t'* axis, whereas the centre of the potential still moves along the *t* axis. This illustrates why the coincidence of the light and the equi-potential spheres (at a particular instant) with respect to one frame of reference does not imply that they ever coincide with respect to another frame of reference. The wavefront of the light pulse is always spherical with respect to both systems of inertial coordinates, whereas the equi-potential surfaces are spherical only with respect to the rest frame of the source.

10

Electrons in One Dimension

In 1964, Little suggested that it might be possible to synthesize a room temperature superconductor using organic materials in which the electrons travelled along certain kinds of chains, effectively confined to one dimension.

The first satisfactory theory of 'ordinary' superconductivity, that of Bardeen, Cooper and Schrieffer (BCS) had appeared a few years earlier, in 1957. The key point was that electrons became bound together in opposite spin pairs, and at sufficiently low temperatures these bound pairs, being boson like, formed a coherent condensate all the pairs had the same total momentum, so all travelled together, a supercurrent. The locking of the electrons into this condensate effectively eliminated the usual single-electron scattering by impurities that degrades ordinary currents in conductors.

But what could bind the electrostatically repelling electrons? The answer turned out to be lattice distortions, as first suggested by Frohlich in 1950. An electron traveling through the crystal attracts the positive ions, the consequent excess of local positive charge attracts another electron. The strength of this binding, and hence the temperature at which the superconducting transition takes place, depends on the rapidity of the lattice response. This was confirmed by the isotope effect: lattice response time obviously depends on the inertia of the lattice, the BCS theory predicted that for a superconducting element with different isotopic varieties, the ratio of the superconducting transition temperatures for pure isotopes T_2 / T_1 was equal to $\sqrt{M_1 / M_2}$, M_1, M_2 being the ion masses, the lighter isotope having the higher transition temperature. This was indeed the case.

Little's idea was that the build up of positive charge by a passing electron could be speeded up dramatically if instead of having to move ions, it need only rearrange other electrons. Unfortunately, there were no obvious three-dimensional candidate materials. However, if the conduction electrons moved along a one-dimensional chain, polarizable side chains might be attached, and rearrangement of the electronic charge distribution in these side chains would respond very rapidly to a passing conduction electron, building up a local positive charge. If this worked, order of magnitude arguments suggested possible enhancement of the transition temperature by a factor $\sqrt{M/m}$ over ordinary superconductors, m being the electron mass.

In the 1970's, various organic materials were synthesized and tested, beginning with one called TTF-TCNQ, in which a set of polymer-like long molecules donated electrons to another set, leaving one-dimensional conductors with partially filled bands, seemingly good candidates for superconductivity. Unfortunately, on cooling these materials surprisingly became *insulators* rather than superconductors! This was the first example of a *Peierls transition*, a widespread phenomenon in quasi one-dimensional systems.

The basic mechanism of the Peierls transition can be understood with a simple model. It is a nice example of applied second-order perturbation theory, including the degenerate case. We examine the model and the result below.

It should be added that in some newer materials the Peierls transition is (unexpectedly) suppressed under high pressure, and superconductivity has in fact been observed in organic salts, but so far only at transition temperatures around one Kelvin, Little's dream is not yet realized.

Second-Order Perturbation Theory

To understand how a one-dimensional conductor might turn into an insulator at low temperatures, we must first become familiar with the simplest model of a one-dimensional conductor:

$$H = H^0 + V = \frac{p^2}{2m} + V(x)$$

with H^0 a gas of noninteracting electrons on a line, and V periodic, that is $V(x+a) = v(x)$, the potential from a line of ions spaced a apart. We'll take the system to have N ions in a total length L, so $L = Na$ and to keep the math simple, we'll require periodic boundary conditions.

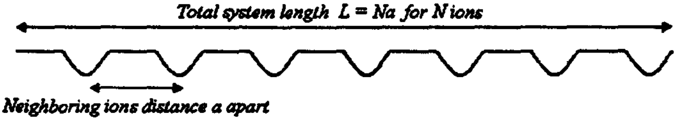

Total system length L = Na for N ions
Neighboring ions distance a apart

Fig. 1 Lonic Potentian seen by electrons in one-dimensional system.

The physics here is that *without* the potential, the electron eigenstates are *plane waves*. The effect of the lattice potential is to partially reflect the waves, like a diffraction grating, generating components at different wavelengths. This effect becomes particularly important when the electron wavelength matches twice the ion spacing. For that case, the reflected and original waves have the same strength, the electron is at a standstill.

The eigenstates of H^0 are then

$$| k\rangle^{(0)} = \frac{1}{\sqrt{L}} e^{i/a}, with e^{i/L} = 1, so\ k = \frac{2\pi n}{L},\ n$$

being an integer. The unperturbed energy eigenvalues, $H^0 \,|\, k_n\rangle^{(0)} = E_n^0 \,|\, k_n\rangle^{(0)}$,

are just $E^0 = \dfrac{\hbar^2 k^2}{2m}$

This is to be understood as

$$H^0 \,|\, k\rangle^{(0)} = E_n^0 \,|\, k_n\rangle^{(0)},$$

$$\text{with } E_n^0 = \frac{\hbar^2 k^2}{2m} \text{ and } k_n = \frac{2\pi n}{L}$$

We are following standard practice here. We shall also write. $\sum_k f(k)$ meaning $\sum_k f(k_n)$. It's worth plotting the (E, k) curve:

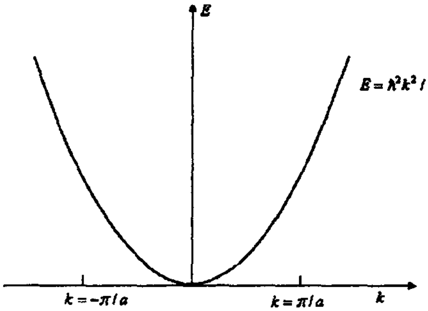


Fig. 2 Energy Mamentum (E,K) Curve for a free Electror in One Cimension.

Suppose we have ions with two electrons each to contribute to this one-dimensional (supposed) conductor. Assuming they move into these plane wave states, in the system ground state they will fill up the lowest energy states up to a maximum *k*-value denoted by $\pm k_F$ (*F* stands for Fermi, this is the *Fermi* momentum.) Where is it?

We know there will be a total of 2*N* electrons. We also know that the allowed values of *k*, from the boundary conditions, are $k_n = 2\pi n / L$, with *n* an integer. In other words, the allowed *k*'s are uniformly spaced $2\pi n / L$ apart, meaning they have a *density* of $L / 2\pi$ in *k*-space, so the total number between $\pm k_F$ is Lk_F / π. The 2*N* electrons will have *N* of each spin, each *k*-state can take two electrons (one of each spin), so $Lk_F / \pi = N = L / a$, and $k_F = \pi / a$.

To do perturbation theory, we must find the matrix elements of *V*(*x*) between eigenstates of H^0:

$$^{(0)}\langle k'|v|k\rangle^{(0)} = \frac{1}{L}\int e^{\imath(k-k')}V(x)\,dx$$

This is just the Fourier component v_{k-k} of *V*(*x*).If *V*(*x*) is periodic with period $a, V_k \neq 0$ only if $k = nk$, *n* an integer, $k = 2\pi/a$. In other words, if a function is periodic with spatial period *a*, the only non-zero Fourier components are those having the same spatial period *a*.

Therefore,

$$V(x) = \sum_n V_{nk^{cinKk}} \text{ and } V_{\cdot nk} = V_{-nK}\ V \text{Since } V(x) \text{ is real;}$$

$K = 2\pi/a.$

The $n = 0$ component of $V(x)$ is of no interest it is just a constant potential, and so can be taken to be zero. Note that this eliminates the trivial first order correction $E_k^1 =^{(0)} \langle k|V|k\rangle^{(0)}$ to the energy eigenvalues.

We shall consider *only* the components $n = +1$ and $n = -1$ of $V(x)$, it turns out that the other components can be treated in similar fashion. For $n = +1$, , the potential only has nonzero matrix elements between the plane wave state k and $k + k$, $k - k$ respectively.

So, the second–order correction to the energy is:

$$E_k^2 = \sum_{k-k'} \left|{}^{(0)}\left(k \,|\, v \,|\, k'\right)^{(0)}\right|^2$$

$$= \frac{\left|{}^{(0)}\langle k \,|\, v \,|\, k+K\rangle^{(0)}\right|^2}{E_K^0 - E_{k-K}^0} + \frac{\left|{}^{(0)}\langle k \,|\, v \,|\, k-K\rangle^{(0)}\right|^2}{E_K^0 - E_{k-K}^0}$$

$$= \frac{|Vk|^2}{E_k^0 - E_{k+K}^0} + \frac{|V-k|^2}{E_k^0 - E_{k-K}^0}$$

This result is reasonable provided the terms are small, that is, the energy differences appearing in the denominators are large compared to the relevant Fourier component V_K. However, this cannot always be true! Notice that the state $k = \pi/a$ has exactly the same unperturbed energy E^0 as the state $= \frac{|Vk|^2}{E_k^0 - E_{k+K}^0} + \frac{|V-k|^2}{E_k^0 - E_{k-K}^0}$: in this case, nondegenerate perturbation theory is clearly wrong. In fact, even for states close to, the energy denominator $E_k^0 - E_{k-k}^0$ is small compared with the numerator $|V_{-k}|^2$, so the series is not converging.

Quasi-degenerate Perturbation Theory Near the Critical Wavelength

The good news is that, despite the many states near $k = \pi / a$ and $k = -\pi / a$ that are close together in energy, for any *one* state k near π / a the potential only has a nonzero matrix element to *one* other state close in energy, the state $k - K$, that is $k = -2\pi / a$. The strategy now is to do what might be called *quasi-degenerate* perturbation theory: to diagonalize the full Hamiltonian in the subspace spanned by these two states $|\,k\rangle^{(0)}, |\,k - k\rangle^{(0)}$. Other states with non-zero matrix elements to these states are relatively much further away in energy, and can be treated using ordinary perturbation theory.

The matrix elements of the full Hamiltonian in the subspace spanned by these two states are:

$$\begin{vmatrix} E_k^0 & V_K^* \\ V_K & E_{k-K}^0 \end{vmatrix}$$

Diagonalizing *within this subspace* gives energy eigenvalues:

$$E_\pm = \frac{1}{2}\left(E_k^0 + E_{k-K}^0\right) \pm \sqrt{\left(\frac{E_k^0 - E_{k-K}^0}{2}\right)^2 + |V_k|^2}$$

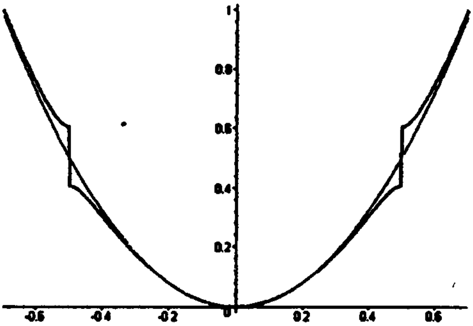


Fig. 3

Notice that, provided

$$\left|E_k^0 - E_{k-k}^0\right| >> |V_k|,$$

to leading order this gives back $E_\pm = E_k^0, E_{k-k}^0$ the order depending on *k*. However, as *k* approaches π/a,

$$\left|E_k^0 - E_{k-k}^0\right|$$

becomes of order, and the energies deviate from the unperturbed values. If *k* is approaching from below, $E_k = E\_ < E_k^0$, and the lower energy is pushed *downwards* by the perturbation: This is a common occurrence with almost degenerate states, perturbations cause the energy levels to 'repel' each other.

For $k = \pi / a, E^0_{k-k} = E^0_{k-2\pi/a} = E^0_k$ At this value of k, the unperturbed states are exactly degenerate, and the perturbation lifts the degeneracy to give $E_{\pm} = E^0_{\pi/a} \pm |V_k|$.

Energy Gaps and Bands

The energy jump, or gap, of $2|V_K|$ at $|k| = p/a$ means that there are no plane wave type eigenstates with energies in that range attempting to integrate Schrodinger's equation in the periodic potential for such an energy gives exponentially growing and decaying solutions. Such energy gaps in fact are present in real crystalline solids, the allowed energies are said to be in 'bands'. The lowest band for our model is from $k = \pi/a$ to π/a Since the allowed values of k are given by $k = 2\pi n/L$, the spacing between adjacent k's is $2\pi / L$ and the total number of k's in the lowest band is $L/a = N$, the same as the number of atoms. Since each electron has two spin states, this implies that a one-dimensional crystal of *divalent* atoms will *just fill* the lowest band with electrons. Therefore, any outside field can only excite an electron to a different state if an energy of at least $2|V_K|$ is supplied for a small electric field, the filled band of electrons will remain in the ground state, there will be no current. This material is an insulator.

On the other hand, if *monovalent* atoms are used, it is clear that the lowest band is *only half full*, adjacent empty electron states are available. The electrons are free to accelerate if an external field is applied. Barring the unexpected, this one-dimensional crystal would be a metal.

Let us now examine how the periodic potential alters the eigenstates. Ignoring the small corrections from plane waves outside the $|k\rangle^{(0)}, |k-K\rangle^{(0)}$ subspace, the eigenstates to this order have the form

$$|k\rangle = a_k \,|k\rangle^{(0)} + a_{k-K}\,|k-K\rangle^{(0)}$$

where

$$\frac{a_{k-K}}{a_k} = \frac{E_- - E^0_k}{V^*_K}$$

from the diagonalization of the 2 × 2 matrix representing the Hamiltonian in the subspace.

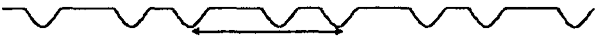
As k increases from 0 towards , the plane wave initially proportional to e^{ikx} has a gradually increasing admixture of $e^{i(k-2x/a)x}$, until at $k = \pi / a$ the two have equal weight meaning that the eigenfunction is now a standing wave. In fact, there are *two* standing wave solutions at $k = \pi / a$, corresponding to the energies below and above the gap. Taking the atoms to have an attractive potential, the lower energy wave has a probability distribution peaking at the atomic positions. The diffractive scattering that gives a left-moving component to a right moving wave is known as Bragg scattering. It also manifests itself in the *group velocity* of the electronic excitations, $v_{group} = d\omega / dk = (1/\hbar) dE / dk$. An electron injected into a one-dimensional metal would not be a plane wave state, but a wavepacket traveling at the group velocity. It is evident that for an injected electron with mean value of k close to π/a, the electron will move very slowly into the metal. This is to be expected the eigenstates become standing waves as $k \to \pi / a$.

For three-dimensional crystals, the situation is far more complicated, but many of the same ideas are relevant. Electron waves are now diffracted by whole planes of atoms, and the three-dimensional momentum space is divided into Brillouin zones, with planes having an energy gap across them.

The Peierls Transition: How Cooling a Conductor can Give an Insulator

As mentioned in the introduction, substances very close to monovalent one-dimensional crystals have been synthesized, and it has been found surprisingly that at low temperatures many of them undergo a transition from metallic to *insulating* behaviour. *What happens is that the atoms in the lattice rearrange slightly, moving from an equally-spaced crystal to one in which the spacing alternates, that is, the atoms form pairs.* This is called *dimerization*, and costs some elastic energy, since for identical atoms the lowest state must be one of equal spacing for any reasonable potential. However, the *electrons* are able to move to a lowe energy state by this manoeuvre.

Total system length $l = Na$ (as before) for N ions



System now has longer period: 2a

Fig. 4 Lonic Potential seen by Electrons in a Dimensional System.

Just how this happens can be understood using the perturbation theory analysis above. For equally spaced atoms, the electrons *half-fill* the band, that is, they fill it up (two electrons, one of each spin, per state) to $|k| = \pi / 2a$.

The crucial point is that if the atoms move together slightly into pairs, *the crystal has a new period* $2a$ *instead of* a. This means that the *potential* now has a nonzero component at $k = -\pi / a$, with a nonzero matrix element between the states $k = \pi / 2a$ and $k = -\pi / 2a$, and so on. From this point, we can rerun the analysis above, except that now the gaps open up at $|k| = \pi / 2a$ instead of at $|k| = \pi / a$.

The important point is that if the electrons fill all the states to $|k| = \pi / 2a$, and none beyond (as would be the case for monovalent atoms) then the opening of a gap at $|k| = \pi / 2a$ means that *all the electrons are in states whose energy is lowered.* To find the *total* energy benefit we need to integrate over k.

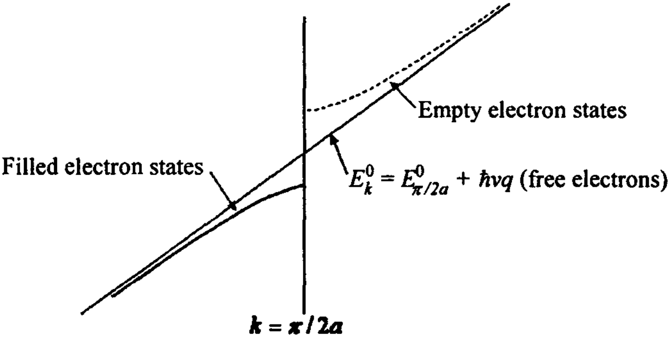


Fig. 5 Change in Electronic Energy levels near $k = \pi / 2a$ from Dimerizatzation: in this region, the Tree Electron curve is approximated with a straight line: $q = k - \pi / 2a$

Calculating the Electronic Energy Gained by Doubling the Lattice Period

It is evident from the above that most of the contribution comes from fairly close to $k = \pi / 2a$ (and of course symmetrically $k = -\pi / 2a$). Since we want to find the total lowering in energy, let us study first the *bare* energy as a function of k, that is, the energy with no potential present. Of course, there isn't much to say: $E_k^0 = \hbar^2 k^2 / 2m$. However, the physics of these one-dimensional systems concerns only excitations near the 'Fermi surface', the boundary between filled (low energy) states at zero temperature and empty states. This 'Fermi surface' is in fact just two points in one dimension: $k = \pm\pi/2a$. In the neighbourhood of these two Fermi points, it is an excellent approximation to replace the gently curving $E_k^0 = \hbar^2 k^2 / 2m$ by *straight line approximations* the slope being *dE/dk = h²k/m=hp/m = hv.*

Linearizing in the neighbourhood of $k = \pi/2a$, then, we take

$$E_k^0 = E_{\pi/2a}^0 + hv\,(k - \pi/2a) = E_{\pi/2a}^0 + hvq.$$

where, $q = k - \pi/2a$

just *k* measured from the Fermi point $\pi/2a$. The variable *q* is *negative* for the relevant states, since they are on the lower energy side. The density of states in *k*-space is a constant $2 \times L/2\pi = L/\pi$, remembering the two spin states per *k*-value. Recall

$$E_\pm = \frac{1}{2}\left(E_k^0 + E_{k-K}^0\right) \pm \sqrt{\left(\frac{E_k^0 - E_{k-K}^0}{2}\right)^2 + |V_K|^2}$$

but now $k = \pi/a$, and the lowering of energy of the electrons (counting it as a positive quantity) is:

$$2\int_0^{\pi/2a} (E_k^0 - E_{k-K})\frac{Ldk}{\pi}$$

$$= \int\left(\frac{1}{2}(E_k^0 - E_{k-K}^0) + \sqrt{\left(\frac{E_k^0 - E_{k-K}^0}{2}\right)^2 + |V_K|^2}\right)\frac{Ldk}{\pi}$$

where the extra factor of 2 counts the symmetrical contribution from the left-hand gap. (In examining the above expression, recall that for the $k > 0$ states we are interested in, $k > \pi/2a, E_k^0 - E_{k-k}^0$ is *negative*. The integrand on the right-hand side is still positive, very small for small *k*, reaching a maximum of $|V_K|$ at $k = \pi/2a$) Putting in our linearized energy approximation,

$$E^{0}_{k-K} = \Delta A \approx \left|V_k\right|^2 \int_{-D}^{-|V_K|} \frac{1}{2hv} \frac{Ldq}{\pi\,|\,q\,|} = \frac{L\left|V_k\right|^2}{hv} 1n \frac{\left|V_k\right|}{D} + hv(k - \pi/2a)$$

$$= E^{0}_{n\ 2a} + hvq,$$

and remembering that now

$$E^{0}_{k-K} = E^{0}_{n\ 2a} - hv\ (k - \pi/2a) = E^{0}_{n\ 2a} - hvq.$$

Since $E^{0}_{x/2a} = E^{0}_{-x/2a}$,

$$\left(E^{0}_{k} - E^{0}_{k-k}\right) = 2v\hbar q$$

Substituting these linearized values in the integral for the total energy lowering:

$$2 \int_{0}^{\pi/2a} (E^{0}_{k} - E_{k-K}) \frac{Ldk}{\pi} = 2 \int_{-D}^{0} \left(v\hbar q + \sqrt{v\hbar q)^2 + |\,V_K|^2}\right) \frac{Ldq}{\pi}$$

where in terms of the variable q we have set the lower limit of integration at $-D$: we can safely be vague about this lower limit, as the integral turns out to be logarithmic.

Since the integral is over negative numbers, and we have taken the positive square root, it is zero for zero V_K, as it must be.

The integral can be done exactly, but it is more illuminating to divide the range of integration into $\left|v\hbar q\right| \leq \left|V_k\right|$ and $\left|v\hbar q\right| > \left|V_k\right|$, then estimate the contributions from these two ranges separately.

First, consider $\left|v\hbar q\right| \leq \left|V_k\right|$. Here the integrand is of order $\left|V_k\right|$, and the region Δq of integration corresponding to $\left|v\hbar q\right| \leq \left|V_k\right|$ is of order so the integral over this range is of order.

Second, in the region $\left|v\hbar q\right| \leq \left|V_k\right|$, we can write

$$2 \int (vhq + \sqrt{(vhq)^2 + \left|V_k\right|^2}) Ldq/\pi$$

$$= 2 \int \left(vhq + \left|vhq\right| \sqrt{1 + \frac{|V_k|2}{(vhq)^2}} \right) Ldq/\pi$$

and expand the square root term. The leading terms cancel since q is negative, and the main contribution comes from the next term. This gives:

$$\Delta A \approx \left|V_k\right|^2 \int_{-D}^{-|Vk|} \frac{1}{2hv} \frac{L dq}{\pi|q|} = \frac{L\left|V_k\right|^2}{hv} \ln \frac{\left|V_k\right|}{D}$$

The important thing here is the logarithm. For sufficiently small $|V_k|$, this large (negative) term will dominate any term which is just proportional to V_k^2. But the elastic energy cost of the lattice 'dimerizing' the atoms forming pairs, so that the distance between atoms alternates on going along the chain must be proportional to V_k^2. This leads to the conclusion that some, probably small, dimerization is always going to happen *a one-dimensional equally spaced chain with one electron per ion is unstable.*