Springer Monographs in Mathematics

Springer
New York
Berlin
Heidelberg
Hong Kong
London
Milan
Paris
Tokyo

Thomas S. Angell Andreas Kirsch

Optimization Methods in Electromagnetic Radiation

With 78 Illustrations

Springer

Thomas S. Angell
Department of Mathematical Sciences
University of Delaware
Newark, DE 19716
USA
angell@math.udel.edu

Andreas Kirsch
Mathematics Institute II
University of Karlsruhe
Englerstr 2.
D-76128 Karlsruhe
Germany
kirsch@math.uni-karlsruhe.de

Mathematics Subject Classification (2000): 78M50, 65K10, 93B99, 47N70, 35Q60, 35J05

Library of Congress Cataloging-in-Publication Data
Angell, Thomas S.
Optimization methods in electromagnetic radiation / Thomas S. Angell, Andreas Kirsch.
p. cm. — (Springer monographs in mathematics)
Includes bibliographical references and index.

1. Maxwell equations—Numerical solutions. 2. Mathematical optimization. 3. Antennas (Electronics)—Design and construction. I. Kirsch, Andreas, 1953– II. Title. III. Series.
QC670.A54 2003
530.14′1—dc22 2003065726

ISBN 978-1-4419-1914-4 ISBN 978-0-387-21827-4 (eBook)
DOI 10.1007/978-0-387-21827-4

© 2004 Springer-Verlag New York, Inc.

Softcover reprint of the hardcover 1st edition 2004

All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher (Springer-Verlag New York, Inc., 175 Fifth Avenue, New York, NY 10010, USA), except for brief excerpts in connection with reviews or scholarly analysis. Use in connection with any form of information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed is forbidden. The use in this publication of trade names, trademarks, service marks, and similar terms, even if they are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights.

9 8 7 6 5 4 3 2 1 SPIN 10951989

Typesetting: Pages created by the authors in LaTeX2e using Springer's SVMono.cls macro.

www.springer-ny.com

Springer-Verlag New York Berlin Heidelberg
A member of BertelsmannSpringer Science+Business Media GmbH

Contents

Preface

The subject of *antenna design*, primarily a discipline within electrical engineering, is devoted to the manipulation of structural elements of and/or the electrical currents present on a physical object capable of supporting such a current. Almost as soon as one begins to look at the subject, it becomes clear that there are interesting *mathematical* problems which need to be addressed, in the first instance, simply for the accurate modelling of the electromagnetic fields produced by an antenna. The description of the electromagnetic fields depends on the physical structure and the background environment in which the device is to operate.

It is the coincidence of a class of practical engineering applications and the application of some interesting mathematical optimization techniques that is the motivation for the present book. For this reason, we have thought it worthwhile to collect some of the problems that have inspired our research in applied mathematics, and to present them in such a way that they may appeal to two different audiences: *mathematicians* who are experts in the theory of mathematical optimization and who are interested in a less familiar and important area of application, and *engineers* who, confronted with problems of increasing sophistication, are interested in seeing a systematic mathematical approach to problems of interest to them. We hope that we have found the right balance to be of interest to both audiences. It is a difficult task.

Our ability to produce these devices at all, most designed for a particular purpose, leads quite soon to a desire to *optimize* the design in various ways. The mathematical problems associated with attempts to optimize performance can become quite sophisticated even for simple physical structures. For example, the goal of choosing antenna feedings, or surface currents, which produce an antenna pattern that matches a desired pattern (the so-called *synthesis problem*) leads to mathematical problems which are *ill-posed* in the sense of Hadamard. The fact that this important problem is not well-posed causes very concrete difficulties for the design engineer.

Moreover, most practitioners know quite well that in any given design problem one is confronted with not only a single measure of antenna performance, but with several, often conflicting, measures in terms of which the designer would like to optimize performance. From the mathematical point of view, such problems lead to the question of multi-criteria optimization whose techniques are not as well known as those associated with the optimization of a single cost functional.

Sooner or later, the question of the efficacy of mathematical analysis, in particular of the optimization problems that we treat in this book, must be addressed. It is our point of view that the results of this analysis is *normative*; that the analysis leads to a description of the theoretically optimal behavior against which the radiative properties of a particular realized design may be measured and in terms of which decisions can be made as to whether that realization is adequate or not.

From the mathematical side, the theory of mathematical optimization, a field whose antecedents pre-date the differential and integral calculus itself, has historically been inspired by practical applications beginning with the apocryphal isoperimetric problem of Dido, continuing with Newton's problem of finding the surface of revolution of minimal drag, and in our days with problems of mathematical programming and of optimal control. And, while the theory of optimization in finite dimensional settings is part of the usual set of mathematical tools available to every engineer, that part of the theory set in infinite dimensional vector spaces, most particularly, those optimization problems whose state equations are partial differential equations, is perhaps not so familiar.

For each of these audiences it may be helpful to cite two recent books in order to place the present one amongst them. It is our view that our monograph fits somewhere between that of Balanis [16] and the recent book of Cessenat [23], our text being more mathematically rigorous than the former and less mathematically intensive than the latter. On the other hand, while our particular collection of examples is not as wide-ranging as in [16], it is significantly more extensive than in [23]. We also mention the book of Stutzman and Thiele [132] which specifically treats antenna design problems exclusively, but not in the same systematic way as we do here. Moreover, to our knowledge the material in our final chapter does not appear outside of the research literature. The recent publications of the IEEE, [35] and [84], while not devoted to the problems of antenna design, are written at a level similar to that found in our book.

While this list of previously published books does not pretend to be complete, we should finally mention the classic work of D.S. Jones [59]. Part of that text discusses antenna problems, including the synthesis problem. The author discusses the approach to the description of radiated fields for wire antennas, and dielectric cylinders, and the integral equation approach to more arbitrarily shaped structures, with an emphasis on methods for the computation of the fields. But while Jones does formulate some of the optimization problems we consider, his treatment is somewhat brief.

The obvious difficulty in attempting to write for a dual audience lies in the necessity to include the information necessary for both groups to understand the basic material. There are few mathematicians who understand the fundamental facts about antennas, or even what is meant by an antenna pattern; it is not unknown but still unusual for engineers to know about ordered vector spaces or even weak-star convergence in Banach spaces.

It is impossible to make this single volume self-contained. Our choice is to present introductory material about antennas, together with some elementary examples in the introductory chapter. That discussion may then serve as a motivation for a more wide-ranging analysis. On the other hand, in order to continue with the flow of ideas, we have chosen to place a summary of the mathematical tools that we will use in the Appendix. That background material may be consulted from time to time as the reader may find necessary and convenient.

The chapter which follows gives some basic information about Maxwell's equations and the asymptotic behavior of solutions which is then used in Chapter 3. There we formulate a general class of optimization problems with radiated fields generated by bounded sources. Most importantly, we give several different measures of antenna performance related to the desired behavior of the radiated fields far from the antenna itself. These cost functionals are related to various properties of this far field and we discuss, in particular, their continuity properties which are of central importance to the problems of optimization.

In the fourth chapter, we concentrate on one particular problem, the synthesis problem mentioned earlier, and on its resolution. Since the problem is ill-posed, we give there a brief discussion of the mathematical nature of this class of problems.

The following two chapters then discuss, respectively, the boundary value problems for the two-dimensional Helmholtz equation, particularly important for treating TE and TM modes, and for the three-dimensional time-harmonic Maxwell equations. Our discussion, in both instances, includes some background in the numerical treatment of those boundary value problems.

Chapter 7, which together with Chapter 8 forms the central part of our presentation, contains the analysis of various optimization problems for specific examples based on the general framework that we constructed in Chapter 3. It is our belief that, while the traditional antenna literature analyzes the various concrete antenna structures somewhat independently, emphasizing the specific properties of each, a more over-arching approach can guide our understanding of the entire class of problems. In any specific application it is inevitable that there will come a time when the very particular details of the physical nature of the antenna will need to addressed in order to complete the design. That being said, the general analytical techniques we study here are applicable to antennas whether they take the form of a planar array of patches or of a line source on the curvilinear surface of the wing of an aircraft. For some of the standard (and simplest) examples, we include a numerical treatment which, quite naturally, will depend on the specifics of the antenna; a curvilinear line source will demand numerical treatment different from an array of radiating dipoles.

In the final chapter, Chapter 8, we address optimization problems arising when (as is most often the case) there is a need to optimize antenna performance with respect to two or more, often conflicting, measures. To give a simple example, there is often a desire to produce both a focused main beam and to minimize the electromagnetic energy trapped close to the antenna itself e.g, to maximize both directivity and gain simultaneously. In such a situation, the end result of such an analysis is a "design curve" which concretely represents the trade-offs that a design engineer must make if the design is to be in some sense optimal.

These problems fall within the general area of multi-criteria optimization **which was initially** investigated in the field of mathematical economics. More recently, such techniques have been applied to structural engineering problems, as for example the problem of the design of a beam with maximal rigidity and minimal mass, and even more recently, in the field of electromagnetics. While there is now an extensive mathematical literature available, the numerical treatment of such problems is most often, but not exclusively, confined to the "bi-criteria" case of two cost functionals. Our numerical illustrations are confined to this simplest case.

We make no pretense that our presentation is complete. Experts in antenna engineering will find many interesting situations have not been discussed. Likewise, experts in mathematical optimization will see that there are many techniques that have not been applied. We will consider our project a success if we can persuade even a few scientists that this general area, lying as it does on the boundary of applied mathematics and engineering, is both an interesting one and a source of fruitful problems for future research.

Finally, we come to the most pleasant of the tasks to face those who write a monograph, namely that of thanking those who have supported and encouraged us while we have been engaged in this task. There are so many!

We should begin by acknowledging the support of the United States Air Force Office of Scientific Research, in particular Dr. Arje Nachman, and the Deutsche Forschungsgemeinschaft for supporting our efforts over several years, including underwriting our continuing research, the writing of this book, the crucial travel between countries, sometimes for only brief periods, sometimes for longer ones.

As well, our respective universities and departments should be given credit for making those visits both possible and comfortable. Without the encouragement of our former and present colleagues, and our research of our research collaborators in particular, the writing of this book would have been impossible.

Specific thanks should be given to Prof. Dr. Rainer Kress of the Institut für Numerische und Angewandte Mathematik, Universität Göttingen, and the late Prof. Ralph E. Kleinman, Unidel Professor of Mathematics at the University of Delaware who introduced us to this interesting field of inquiry.

1

Arrays of Point and Line Sources, and Optimization

1.1 The Problem of Antenna Optimization

Antennas, which are devices for transmitting or receiving electromagnetic energy, can take on a variety of physical forms. They can be as simple as a single radiating dipole, or far more complicated structures consisting, for example, of nets of wires, two-dimensional patches of various geometric shapes, or solid conducting surfaces. Regardless of the particular nature of the device, the goal is always to transmit or receive electromagnetic signals in a desirable and efficient manner. For example, an antenna designed for use in aircraft landing often is required to transmit a signal which is contained in a narrow horizontal band but a wide vertical one.

This example illustrates a typical problem in antenna design in which it is required to determine an appropriate "feeding" of a given antenna structure in order to obtain a desired radiation pattern far away from the physical antenna. We will see, as we proceed with the theory and applications in later chapters, that a number of issues are involved in the design of antennas intended for various purposes. Moreover, these issues are amenable to *systematic* mathematical treatment when placed in a suitably general framework.

We will devote the next chapter to a discussion of Maxwell's Equations and Chapter 3 to the formulation and general framework for treating the optimization problems. We begin with specific applications in Chapter 4 in which we analyse the synthesis problem whose object is to feed a particular antenna so that, to the extend possible, a prescribed radiation pattern is established. Chapters 5 and 6 discuss the underlying two and three dimensional boundary value problems, and subsequent chapters are devoted to the analysis of various optimization problems associated with the design and control of antennas. In this first chapter we introduce the subject by discussing, on a somewhat *ad hoc* basis, what is perhaps the most extensively studied class of antennas: arrays of elementary radiators and one-dimensional sources.

We make no pretense of completeness; we do not intend to present an exhaustive treatment of what is known about these antennas, even were that possible. There are many books on the subject of linear arrays alone, and the interested reader may consult the bibliography for some of the more recent treatises. Our purpose here, and in subsequent chapters, is to present a single *mathematical framework* within which a large number of antenna problems may be set and effectively treated.

Roughly speaking, this framework consists of a mathematical description of the relation between the electromagnetic currents fed to an antenna and the resulting radiated field. Of particular interest will be the "far field" which describes the radiated field at large distances (measured in terms of wave lengths and the geometry of the antenna), as well as certain measures of antenna "efficiency" or "desirability". Such measures are often expressed in terms of the proportion of input power radiated into the far field in the first case, or in terms of properties of the far field itself in the second. In addition, there are always constraints of various kinds which must be imposed if the design is to be practical e.g., the desired pattern must be attained with limited power input, or the radiation outside a given sector must meet certain bounds. The problems we treat here therefore fall into the category of *constrained optimization problems.*

We set the stage by looking at two specific problems, the problem of optimizing directivity and efficiency factors of linear and circular arrays and line sources, and the "Dolph-Tschebysheff" problem which is concerned with optimizing the relationship between beam-width and side lobe level. We will return to various versions of these problems in later chapters. We begin by reviewing some basic facts about simple sources, which we will derive rigorously later. Once we have these facts at hand, we discuss optimization problems and some methods for their resolution.

1.2 Arrays of Point Sources

By an array of point sources we mean an antenna consisting of several individual and distinguishable dipole elements whose centers are finitely separated. For a **linear** or **circular** array they are assumed to lie on a straight line or a circle, respectively. In Chapter 2, Section 2.10, we will show that the form of the electric field generated by a set of $2N+1$ electric dipoles with arbitrary locations $\boldsymbol{y}_n \in \mathbb{R}^3$, $n=-N,\dots,N$, with (common) moments $\hat{\boldsymbol{p}}$ is

$$\boldsymbol{E}(\boldsymbol{x}) \;=\; \frac{i\omega\mu_0}{4\pi}\,\frac{\mathrm{e}^{ikr}}{r}\big[\hat{\boldsymbol{x}}\times(\hat{\boldsymbol{p}}\times\hat{\boldsymbol{x}})\big]\sum_{n=-N}^{N} a_n\,\mathrm{e}^{-ik\boldsymbol{y}_n\cdot\hat{\boldsymbol{x}}} \;+\; \mathcal{O}\left(\frac{1}{r^2}\right). \qquad (1.1)$$

where we have used spherical coordinates (r,ϕ,θ). Here, k is the **wave number** which is related to the **wave length** λ by $k=2\pi/\lambda$. The complex number a_n is the excitation coefficient for the n–th element, and $\hat{\boldsymbol{x}} = \boldsymbol{x}/\left|\boldsymbol{x}\right| =$

$(\sin\theta\,\cos\phi\,,\,\sin\theta\,\sin\phi\,,\,\cos\theta)^\top \in S^2$ is the unit vector in the direction of the radiated field[1]. Once the direction $\hat{\boldsymbol{p}}$ of the dipole orientation is fixed, the far field of $\boldsymbol{E}$ is entirely determined by the **array factor**, $f(\hat{\boldsymbol{x}})$, defined as

$$f(\hat{\boldsymbol{x}}) \;=\; f(\theta,\phi) \;=\; \sum_{n=-N}^{N} a_n\, \mathrm{e}^{-ik\boldsymbol{y}_n\cdot\hat{\boldsymbol{x}}}, \quad \hat{\boldsymbol{x}} \in S^2. \tag{1.2}$$

We note that the array factor should be distinguished from the far field pattern

$$\boldsymbol{E}_\infty(\hat{\boldsymbol{x}}) \;=\; \hat{\boldsymbol{x}} \times (\hat{\boldsymbol{p}} \times \hat{\boldsymbol{x}}) \sum_{n=-N}^{N} a_n\, \mathrm{e}^{-ik\boldsymbol{y}_n\cdot\hat{\boldsymbol{x}}}, \quad \hat{\boldsymbol{x}} \in S^2\,. \tag{1.3}$$

This is not only because $\boldsymbol{E}_\infty$ is a vector field and f a scalar quantity but also because the magnitudes differ by the factor $\big|\hat{\boldsymbol{x}} \times (\hat{\boldsymbol{p}} \times \hat{\boldsymbol{x}})\big|$. In spherical coordinates (ϕ,θ) we have for $\hat{\boldsymbol{p}} = \hat{\boldsymbol{e}}_3$ that $\big|\hat{\boldsymbol{x}} \times (\hat{\boldsymbol{p}} \times \hat{\boldsymbol{x}})\big| = \sin\theta$, $0 \le \theta \le \pi$.

We now specify a particular direction $\hat{\boldsymbol{x}}_0 \in S^2$ which we will keep fixed during the following discussion. We think of $\hat{\boldsymbol{x}}_0$ as that direction in which we would like to maximize the power of the array factor. Then it is convenient to rewrite (1.2) in the form

$$f(\hat{\boldsymbol{x}}) \;=\; f(\theta,\phi) \;=\; \sum_{n=-N}^{N} a_n\, \mathrm{e}^{-ik\boldsymbol{y}_n\cdot(\hat{\boldsymbol{x}}-\hat{\boldsymbol{x}}_0)}, \quad \hat{\boldsymbol{x}} \in S^2\,, \tag{1.4}$$

where we have replaced a_n by $a_n \exp(ik\boldsymbol{y}_n \cdot \hat{\boldsymbol{x}}_0)$ which is only a change in the phase of the complex number a_n. From this form we see directly that $|f(\hat{\boldsymbol{x}})| \le \sum_{n=-N}^{N} |a_n|$ for all $\hat{\boldsymbol{x}}$ and $f(\hat{\boldsymbol{x}}_0) = \sum_{n=-N}^{N} a_n$. Therefore, if all coefficients a_n are in phase (i.e. if there exists some $\delta \in [0, 2\pi]$ with $a_n = |a_n| \exp(i\delta)$ for all n) then from (1.4), $|f(\hat{\boldsymbol{x}})|$ attains its maximal value at $\hat{\boldsymbol{x}} = \hat{\boldsymbol{x}}_0$.

1.2.1 The Linear Array

Let us first consider the simplest case of a linear array of uniformly spaced elements which we assume to be located symmetrically along the x_3-axis of a three dimensional Cartesian coordinate system. The locations are thus given by $\boldsymbol{y}_n = nd\,\hat{\boldsymbol{e}}_3$, $n = -N, \ldots, N$, with inter-element spacing d. The array factor reduces to

$$f(\theta) \;=\; \sum_{n=-N}^{N} a_n\, \mathrm{e}^{-inkd(\cos\theta-\cos\theta_0)}\,, \quad 0 \le \theta \le \pi\,, \tag{1.5}$$

where $\theta_0 \in [0,\pi]$ corresponds to $\hat{\boldsymbol{x}}_0$. An array with $\theta_0 = \pi/2$ is called a **broadside array** since the main beam is perpendicular to the axis of the

[1] By S^{d-1} we denote the unit sphere in $\mathbb{R}^d$. Thus in $\mathbb{R}^2$, S^1 is the unit circle.

antenna while the values $\theta_0 = 0$ or $\theta_0 = \pi$ correspond to **end-fire arrays** since the main beams are in the same direction as the axis of the array.

An array which is fed by the constant coefficients

$$a_n = \frac{1}{2N+1}, \quad n = -N, \dots, N, \tag{1.6}$$

is called a **uniform array**. With respect to the original form (1.2) the coefficients $a_n = \exp(inkd\cos\theta_0)/(2N+1)$ have constant magnitude and linear phase progression. In this case, the array factor is given by

$$f(\theta) = \frac{1}{2N+1}\sum_{n=-N}^{N} \mathrm{e}^{-inkd(\cos\theta-\cos\theta_0)} = \frac{1}{2N+1}\sum_{n=-N}^{N} \mathrm{e}^{-in\gamma},$$

where we have introduced the auxiliary variable $\gamma = \gamma(\theta, \theta_0) = kd(\cos\theta - \cos\theta_0)$. The following simple calculation shows how to rewrite f in the form (setting $z := \exp(-i\gamma)$):

$$\begin{aligned} f(\theta) = \frac{1}{2N+1}\sum_{n=-N}^{N} z^n &= \left(\frac{1}{2N+1}\right)\frac{z^{N+1} - z^{-N}}{z-1} \\ = \left(\frac{1}{2N+1}\right)\frac{z^{N+\frac{1}{2}} - z^{-(N+\frac{1}{2})}}{z^{\frac{1}{2}} - z^{-\frac{1}{2}}} &= \frac{\sin(N+\frac{1}{2})\gamma}{(2N+1)\sin\frac{\gamma}{2}}, \end{aligned}$$

so that

$$f(\theta) = \frac{\sin(2N+1)\frac{\gamma}{2}}{(2N+1)\sin\frac{\gamma}{2}} = \frac{\sin\left[\frac{2N+1}{2}kd(\cos\theta-\cos\theta_0)\right]}{(2N+1)\sin\left[\frac{kd}{2}(\cos\theta-\cos\theta_0)\right]}. \tag{1.7}$$

A typical graph for $\left|\frac{\sin[(2N+1)\gamma/2]}{(2N+1)\sin(\gamma/2)}\right|$ as a function of γ then looks like the curve in Figure 1.1.

From the equation (1.7) we see some of the main features of uniform arrays. Besides the **main lobe** centered at $\theta = \theta_0$, i.e. $\gamma = 0$, we observe a number of **side lobes** of the same magnitude at locations $\gamma = 2m\pi$, $m \in \mathbb{Z}$, $m \neq 0$. These are called **grating lobes**. Returning to the definition of γ, as θ varies between 0 and π, the variable $\gamma = kd(\cos\theta - \cos\theta_0)$ varies over an interval of length $2kd$ centered at $\gamma_0 = -kd\cos\theta_0$. This interval is called the **visible range**. Its length depends on d while its position depends on θ_0. In particular, for the broadside array the visible range is $[-kd, kd]$ while for the end-fire array it is either $[-2kd, 0]$ or $[0, 2kd]$. We note that for the uniform array the grating lobes lie outside the visible range provided $kd < 2\pi$ and $kd < \pi$ for a broadside and an end-fire array, respectively.

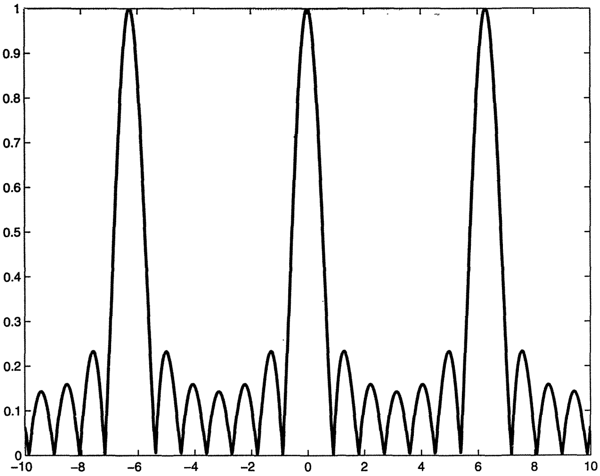


Fig. 1.1. $\gamma \mapsto \left|\frac{\sin[(2N+1)\gamma/2]}{(2N+1)\sin(\gamma/2)}\right|$ for $N = 3$ (the seven element array)

From our expression (1.7) for $|f(\theta)|$ and its graph, we notice certain further typical features. The graph is oscillatory and the zeros (or **nulls**) which define the extent of the individual lobes correspond to the roots of the equations

$$\frac{2N+1}{2}\,kd\,(\cos\theta - \cos\theta_0) \;=\; j\pi, \quad j = \pm 1, \pm 2, \ldots \tag{1.8}$$

The angular separation between the first nulls on each side of the main beam can be approximated for large N by a simple use of Taylor's theorem. Indeed, the condition for the first null corresponding to $j = -1$ is

$$kd\,(\cos\theta_1 - \cos\theta_0) \;=\; \frac{-2\pi}{2N+1}\,. \tag{1.9}$$

The difference on the left can be estimated, for large N, using the **wave length** $\lambda = \frac{2\pi}{k}$, by

$$\begin{aligned} -\frac{\lambda}{(2N+1)d} &= \cos\theta_1 \;-\; \cos\theta_0 \\ &\simeq -(\theta_1 - \theta_0)\,\sin\theta_0 \;-\; \frac{(\theta_1 - \theta_0)^2}{2}\,\cos\theta_0\,. \end{aligned}$$

Thus, for a broadside array (i.e. $\theta_0 = \pi/2$), the angular separation is $\frac{2\lambda}{(2N+1)d}$ while the corresponding result for an end-fire array (i.e. $\theta_0 = 0$) is $2\left(\frac{2\lambda}{(2N+1)d}\right)^{1/2}$. Comparison of these results shows that, for large N, the beam-width for a broadside array is smaller than that for an end-fire array. By **beam-width** of the main lobe we mean just the angular separation between the first nulls on each side. Moreover, since the nulls in the broadside case are given by

$$\theta_j = \arccos\left(\frac{j\lambda}{(2N+1)d}\right), \quad j = \pm 1, \pm 2, \ldots, \tag{1.10}$$

for positive j we must have $0 \leq j\lambda/(2N+1)d \leq 1$ or $j\lambda \leq (2N+1)d$. It follows from this last inequality that such an array has $(2N+1)d/\lambda$ nulls on *each* side of the main lobe so that, if $d = \lambda/2$, there are $2N$ nulls since $2N+1$ is odd.

The fact that the beam-width of the main lobe varies inversely with the size of the array suggests that a narrow beam-width can be obtained simply by increasing the number of elements in the array. The expression for the nulls shows, however, that the number of side lobes likewise increases with N, see Figure 1.2. Since the occurrence of these side lobes indicates that a considerable part of the radiated energy is distributed in unwanted directions, it should be clear that there is a trade-off between narrowing the main beam, and increasing the number of side lobes. We will come back to this idea of a "tradeoff" later in this chapter and again in Chapter 8.

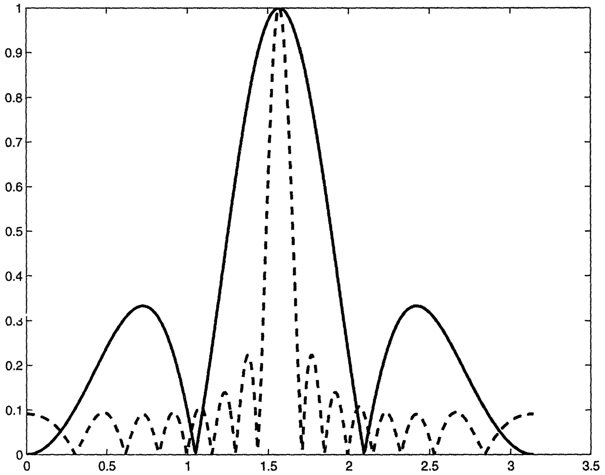


Fig. 1.2. Arrays for 3 and 11 Element Arrays ($\lambda/d = 1.5$)

It is also possible to keep the number of sources fixed, and then to study the dependence of the array pattern on the spacing d. Here again, we see that an increase in the spacing, while narrowing the main beam, increases the number of side lobes. In both cases then, the narrowing of the main beam is made at the expense of the power radiated into that angular sector (see Figure 1.3).

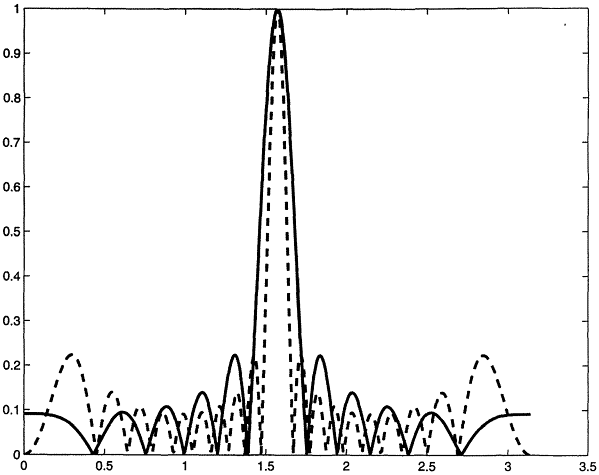


Fig. 1.3. Effect of Increasing Spacing ($N = 5$): $\lambda/d = 2$ (solid) and $\lambda/d = 1.1$ (dashed)

The specification of the pattern is given sometimes not only by the beamwidth of the main lobe, but also by the ratio ρ between the maximum value of the main lobe and that of the largest side lobe which is often, but not always, the first side lobe. It is therefore of interest to be able to compute the various maxima of the array factor.

Clearly, these local maxima occur when $\frac{d}{d\theta}|f(\theta)^2| = 0$ (and $f(\theta) \neq 0$). In the present case, that of a uniform array, a simple computation shows these critical points occur at solutions of the transcendental equation

$$\tan\left[\frac{2N+1}{2}kd\sin\theta\right] = (2N+1)\tan\left[\frac{kd}{2}\sin\theta\right].$$

Thus, the points where maxima occur, as well as the maximal values themselves, can be determined numerically.

While these derivations depend on the representation of the far field pattern in the form (1.7) which assumes that the feeding is uniform, we could imagine choosing different, *non-uniform feedings*. We expect that a different choice of weights would lead to alterations in the far field pattern. Indeed, a typical problem of design is to feed the antenna in such a way that the prominent main beam contains most of the power, while the side lobes, which represent undesirable power loss, are negligible. For example, we may allow feeding coefficients in (1.5) other than the constant ones $a_n = 1/(2N+1)$, $n = -N, \dots, N$, in an attempt to suppress the unwanted side lobes. We illustrate this possibility by considering two feeding distributions which are called, respectively, **triangular** and **binomial**. If the coefficients appearing in the expression (1.5) for the array pattern are symmetric (i.e. $a_{-n} = a_n$) then we can write the array pattern in the form

$$f(\theta) \;=\; a_0 \;+\; 2\sum_{n=1}^{N} a_n \cos\bigl(n\gamma(\theta)\bigr) \quad \text{where } \gamma(\theta) = kd(\cos\theta - \cos\theta_0)\,. \quad (1.11)$$

In order to see concretely the effects of using these non-uniform distributions, let us consider a seven element broadside array (i.e. $\theta_0 = \pi/2$) in which the separation of the elements is $d = \lambda/2$. With this spacing, the parameter $\gamma(\theta) = \pi\cos\theta$. The triangular distribution for this case has coefficients $a_n = 4 - n$, $n = 0, \dots, 3$ while the binomial feeding is defined by the coefficients $a_n = \binom{6}{3-n} = \frac{6!}{(3-n)!\,(3+n)!}$, $n = 0, 1, 2, 3$. Figures 1.4, 1.5, and 1.6 compare these two tapered distributions with the array factor for a uniformly fed seven element broadside array (as a function of θ).

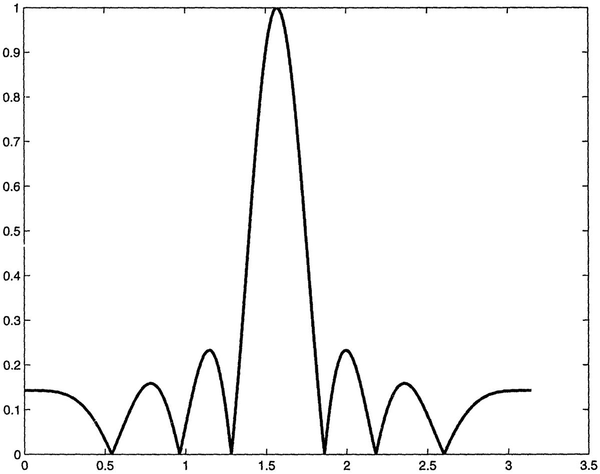


Fig. 1.4. Array for Uniform Feeding (Broadside Array, $N = 3$)

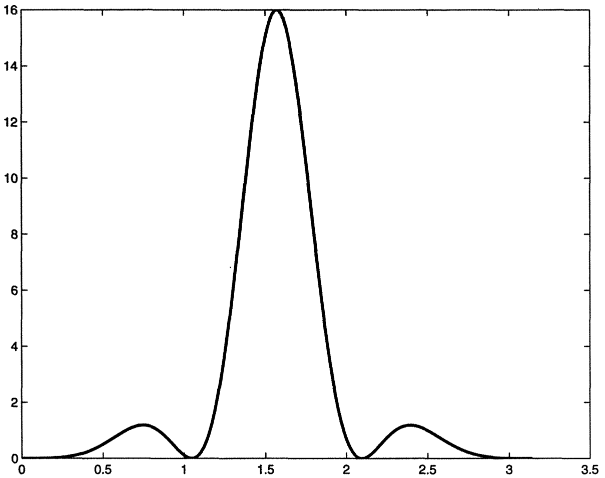


Fig. 1.5. Array for Triangular Feeding (Broadside Array, $N = 3$)

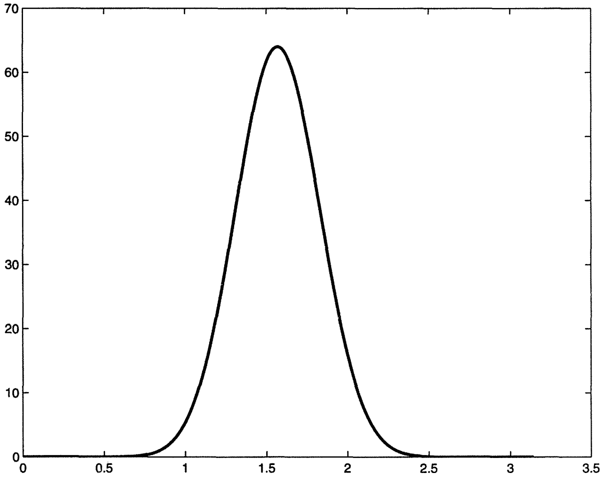


Fig. 1.6. Array for Binomial Feeding (Broadside Array, $N = 3$)

It is evident that, while the triangular distribution partially suppresses the side lobes, the binomial distribution does so completely. One might conclude that, since side lobes are undesirable features of an array pattern, the binomial distribution is in some sense optimal. However, numerical approximation of the first nulls lead to beam-widths of approximately 1.86, 2.09, and 3.14 respectively so that it is again clear that the suppression of the side lobes comes at the expense of beam-width.

The question that we are confronted with is how such a trade-off is to be evaluated. One way to do this is to introduce the notion of the *directivity* of an antenna; we turn to this idea in Section 1.3. But first we analyse a configuration other than a linear array.

1.2.2 Circular Arrays

In this subsection, we will consider a second example of an array, which has found applications in radio direction finding, radar, and sonar: the circular array. Our discussion will be parallel to that of the linear case but will be somewhat abbreviated since many of the ideas that we will meet have analogs in the linear case and are now familiar.

Our object is to analyse a single circular array consisting of N elements equally spaced on the circumference of a circle of radius a which we take to lie in the (x, y)-plane and to have center at the origin. If we measure the phase excitation relative to the center of the circle (at which an element may or may not be present), the mth element has the position vector

$$\boldsymbol{y}_m \;=\; a\cos\phi_m\,\hat{\boldsymbol{e}}_1 \;+\; a\sin\phi_m\,\hat{\boldsymbol{e}}_2 \;=\; a\left(\cos\phi_m\,,\,\sin\phi_m\,,\,0\right)^\top$$

where

$$\phi_m \;=\; \frac{2\pi}{N}\,m\,, \qquad m = 1,\dots,N.$$

With this notation, the array factor for the circular array becomes (compare equation (1.4))

$$f(\hat{\boldsymbol{x}}) \;=\; f(\theta,\phi) \;=\; \sum_{m=1}^{N} a_m\,\mathrm{e}^{-ik\boldsymbol{y}_m\cdot(\hat{\boldsymbol{x}}-\hat{\boldsymbol{x}}_0)} \;=\; \sum_{m=1}^{N} a_m\,\mathrm{e}^{-ik\boldsymbol{y}_m\cdot\boldsymbol{z}} \tag{1.12}$$

where $\boldsymbol{z}$ is the projection of $\hat{\boldsymbol{x}} - \hat{\boldsymbol{x}}_0$ onto the plane of the array, i.e.

$$\boldsymbol{z} \;=\; \boldsymbol{z}(\theta,\phi) \;=\; \left(\sin\theta\cos\phi - \sin\theta_0\cos\phi_0\,,\,\sin\theta\sin\phi - \sin\theta_0\sin\phi_0\,,\,0\right)^\top,$$

(θ_0, ϕ_0) denoting the spherical coordinates of $\hat{\boldsymbol{x}}_0$. Introducing new variables ρ and ξ to be the plane polar coordinates of $\boldsymbol{z}$, i.e. $\boldsymbol{z} = \rho\,(\cos\xi, \sin\xi, 0)^\top$, yields

$$f(\theta,\phi) \;=\; \sum_{m=1}^{N} a_m\,\mathrm{e}^{-ika\rho\cos(\xi-\phi_m)} \tag{1.13}$$

where the dependence on θ and ϕ is through ρ and ξ. Comparison of this form with the expression (1.5) shows that now the array factor is a function of both ϕ and θ.

In the case of *constant feeding* $a_m = \frac{1}{N}$, $m = 1, \ldots, N$, we use the Jacobi-Anger expansion in terms of Bessel functions (cf. [30])

$$e^{-iz\cos t} = \sum_{n=-\infty}^{\infty} J_n(z)\, e^{in(t-\pi/2)}\,, \tag{1.14}$$

to arrive at

$$\begin{aligned} f(\theta,\phi) &= \frac{1}{N}\sum_{m=1}^{N} e^{-ika\rho\cos(\xi-\phi_m)} \\ &= \frac{1}{N}\sum_{n=-\infty}^{\infty} J_n(ka\rho)\, e^{in(\xi-\pi/2)} \sum_{m=1}^{N} e^{-inm\frac{2\pi}{N}} \\ &= \sum_{n=-\infty}^{\infty} J_{nN}(ka\rho)\, e^{inN(\xi-\pi/2)} \\ &= J_0(ka\rho) \;+\; 2\sum_{n=1}^{\infty}(-i)^{nN}\, J_{nN}(ka\rho)\, \cos(nN\xi) \end{aligned} \tag{1.15}$$

where nN is the product of the running index n and the total number of elements N. In the derivation, we have used the identity for the Bessel functions J_m, namely that $J_{-m}(z) = (-1)^m J_m(z)$. The term with the zeroth-order Bessel function $J_0(ka\rho)$ is called the *principal term*; the rest of the terms may be viewed as perturbations. Indeed, from the asymptotic behaviour of the Bessel functions (cf. [50])

$$J_m(t) = \frac{t^m}{2^m\, m!}\left[1 + \mathcal{O}(1/m)\right] \quad \text{as } m \to \infty \tag{1.16}$$

uniformly for t in compact subsets of $[0,\infty)$ and the estimate $(nN)! \geq N!\,[(n-1)N]!$ for $n, N \geq 1$, we note that

$$\begin{aligned} |f(\theta,\phi) - J_0(ka\rho)| &\leq c\sum_{n=1}^{\infty}\frac{(ka\rho)^{nN}}{2^{nN}\,(nN)!} \;\leq\; c\left(\frac{ka\rho}{2}\right)^N \frac{1}{N!}\sum_{n=0}^{\infty}\frac{(ka\rho)^{nN}}{2^{nN}\,(nN)!} \\ &\leq c\left(\frac{ka\rho}{2}\right)^N \frac{1}{N!}\, e^{ka\rho/2}\,. \end{aligned}$$

Therefore, for large N, the pattern $f(\theta,\phi)$ is well approximated by $J_0(ka\rho) = J_0(ka\,|\boldsymbol{z}|)$.

There is a slightly different interpretation of this fact. We can consider the array factor $f(\theta,\phi)$ from (1.13), in the case of $a_m = 1/N$, as the discretization of the continuous circular line source of radius a

$$\tilde{f}(\theta,\phi) = \frac{1}{2\pi}\int_0^{2\pi} e^{-ika\rho\cos(\xi-s)}\, ds\,. \tag{1.17}$$

The application of the Jacobi-Anger expansion (1.14) in this expression also yields $\tilde{f}(\theta,\phi) = J_0(ka\rho)$.

As a particular example we consider first the *broadside* case $\theta_0 = 0$ (or $\theta_0 = \pi$), i.e. where the desirable beam is perpendicular to the plane of the array. The vector $\boldsymbol{z}$ takes the form $\boldsymbol{z} = \sin\theta\,(\cos\phi, \sin\phi, 0)^\top$ and thus $\rho = \sin\theta$ and $\xi = \phi$. This gives the approximate far field $\tilde{f}(\theta,\phi) = J_0(ka\sin\theta)$ which is **omnidirectional**, i.e. independent of ϕ, see Figure 1.7 for $a = \lambda/2$, i.e. $ka = \pi$.

In the case $\theta_0 = \pi/2$, where the beam is in the plane of the array, we have $\boldsymbol{z} = (\sin\theta\cos\phi - \cos\phi_0, \sin\theta\sin\phi - \sin\phi_0, 0)^\top$. Here we find both horizontal (azimuthal) patterns which lie in the plane $\theta = \pi/2$ of the array and vertical patterns which lie in the vertical plane corresponding to $\sin(\phi - \phi_0) = 0$. For the horizontal pattern we have, after some elementary calculations,

$$\begin{aligned}\boldsymbol{z} &= (\,\cos\phi - \cos\phi_0\,,\, \sin\phi - \sin\phi_0\,,\, 0)^\top \\ &= 2\sin\frac{\phi-\phi_0}{2}\,\Big(\cos\frac{\phi+\phi_0+\pi}{2}\,,\, \cos\frac{\phi+\phi_0+\pi}{2}\,,\, 0\Big)^\top ,\end{aligned}$$

and so

$$\tilde{f}(\pi/2,\phi) \;=\; J_0\big(2ka\sin\frac{\phi-\phi_0}{2}\big)\,.$$

For the vertical pattern corresponding to $\phi = \phi_0$ or $\phi = \phi_0 + \pi$ we have $\boldsymbol{z} = (\sin\theta - 1)\,\big(\cos\phi_0, \sin\phi_0, 0\big)^\top$ or $\boldsymbol{z} = -(\sin\theta + 1)\,\big(\cos\phi_0, \sin\phi_0, 0\big)^\top$, respectively, and thus

$$\tilde{f}(\theta,\phi_0) \;=\; J_0\big(ka(1-\sin\theta)\big)\,, \quad \tilde{f}(\theta,\phi_0+\pi) \;=\; J_0\big(ka(1+\sin\theta)\big)\,,$$

respectively. Plots of these patterns $|\tilde{f}|$ for $a = \lambda/2$, i.e. $ka = \pi$, and $\phi_0 = 0$ are given in Figures 1.8 and 1.9 below.

A convenient form of representing the array patterns as well as some other quantities we will derive from it as, for example, directivity is to use the notations from vector analysis. We denote by

$$(\boldsymbol{a},\boldsymbol{b}) \;=\; \sum_{j=1}^{m} a_j\,\overline{b_j} \quad\text{and}\quad |\boldsymbol{a}| \;=\; \sqrt{(\boldsymbol{a},\boldsymbol{a})} \;=\; \sqrt{\sum_{j=1}^{m} |a_j|^2}\,,$$

the inner product and Euclidean norm, respectively, in $\mathbb{C}^m$. If the point sources are located at $\boldsymbol{y}_n$, $n = -N, \dots, N$, we introduce a vector $\boldsymbol{a} \in \mathbb{C}^{2N+1}$ whose components are the feeding coefficients a_n, and the vector

$$\boldsymbol{e} \;=\; \boldsymbol{e}(\hat{\boldsymbol{x}}) \;=\; \Big(\mathrm{e}^{ik\boldsymbol{y}_{-N}\cdot(\hat{\boldsymbol{x}}-\hat{\boldsymbol{x}}_0)}\,,\,\dots\,,\,\mathrm{e}^{ik\boldsymbol{y}_{N}\cdot(\hat{\boldsymbol{x}}-\hat{\boldsymbol{x}}_0)}\Big)^\top \;\in\; \mathbb{C}^{2N+1}$$

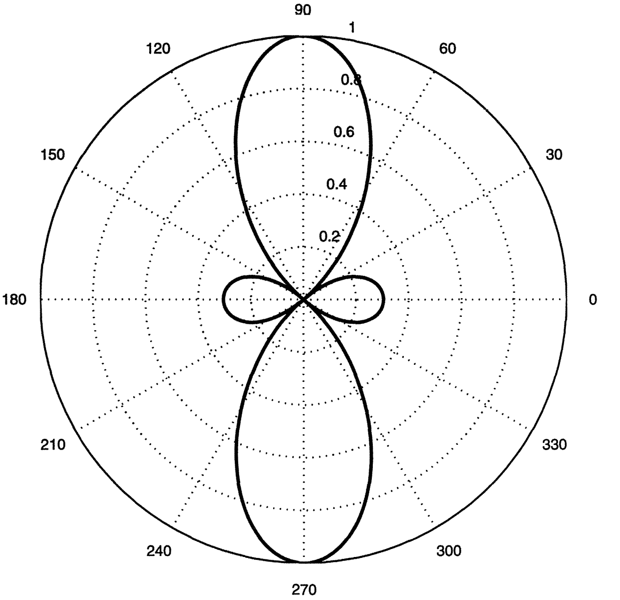


Fig. 1.7. $\theta \mapsto |\tilde{f}(\theta, \phi)|$ for $\theta_0 = 0$ in the (x, z)–plane

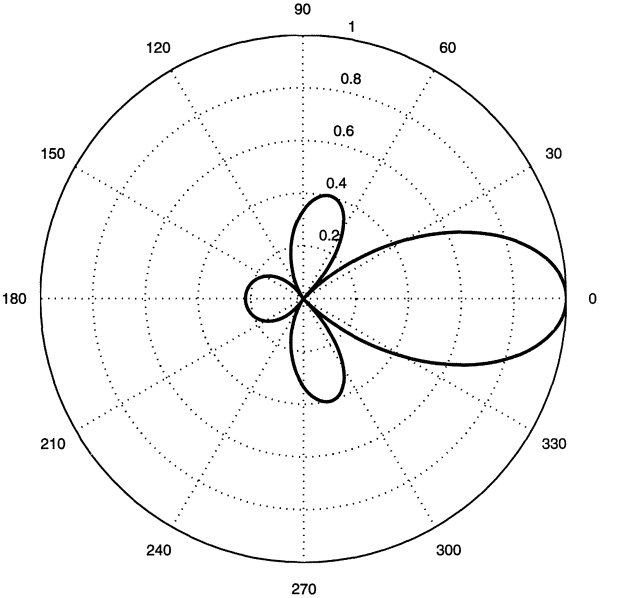


Fig. 1.8. $\phi \mapsto |\tilde{f}(\pi/2, \phi)|$ for $\theta_0 = \pi/2$ and $\phi_0 = 0$ in the (x, y)–plane

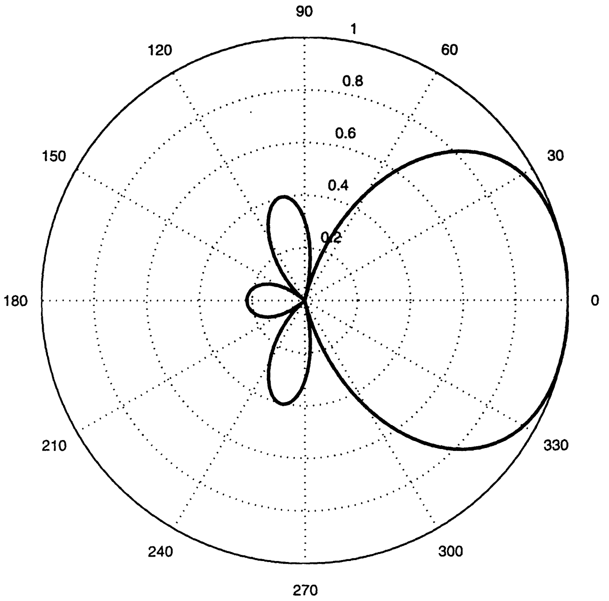


Fig. 1.9. $\theta \mapsto |\tilde{f}(\theta,\phi)|$ for $\theta_0 = \pi/2$ and $\phi = \phi_0$ or $\phi = \phi_0 + \pi$ in the (x,z)−plane

(note that we write vectors in column-form). Then the array factor may be represented simply as the complex inner product

$$f(\hat{\boldsymbol{x}}) \;=\; \sum_{n=-N}^{N} a_n \, \mathrm{e}^{-ik\boldsymbol{y}_n\cdot(\hat{\boldsymbol{x}}-\hat{\boldsymbol{x}}_0)} \;=\; (\boldsymbol{a}, \boldsymbol{e}(\hat{\boldsymbol{x}}))\,. \tag{1.18}$$

In order to compare this form of the array pattern with other feeding mechanisms which will be introduced later it is convenient to consider the operator

$$\mathcal{K} : \mathbb{C}^{2N+1} \longrightarrow C(S^2)\,, \quad \boldsymbol{a} \mapsto f = (\boldsymbol{a}, \boldsymbol{e})\,. \tag{1.19}$$

This linear operator maps the finite dimensional space $\mathbb{C}^{2N+1}$ of feeding coefficients into the space $C(S^2)$ of continuous functions on the unit sphere. It is one-to-one since $\mathcal{K}\boldsymbol{a} = 0$ implies that $\sum_{n=-N}^{N} a_n \, \mathrm{e}^{-ik\boldsymbol{y}_n\cdot(\hat{\boldsymbol{x}}-\hat{\boldsymbol{x}}_0)} = 0$ for all $\hat{\boldsymbol{x}} \in S^2$ and thus $a_n = 0$ for all n.

1.3 Maximization of Directivity and Super-gain

The discussion in the preceeding section has shown that the behavior of the radiated far field pattern of a source depends on the "feeding" or currents on the radiating elements. The ability to change those currents affords us the possibility of manipulating the radiated pattern in the far field and, moreover, the possibility of doing so in an "optimal" fashion. In order to define what is an optimal pattern however, we must have some measure of desirability. It is to this question that we devote the first subsection. In part 1.3.2 we turn to the optimization problems.

1.3.1 Directivity and Other Measures of Performance

Measures of antenna performance are scalar quantities which, in some way measure desirable properties of the antenna pattern as a functional of the inputs to the antenna and, perhaps, other parameters of interest as, for example, inter-element spacing. In keeping with the introductory nature of the present chapter, we will discuss some traditional measures, in particular the *directivity* of an antenna and the *signal-to-noise ratio.* A more comprehensive discussion of performance measures will be deferred until Chapter 3. When treating arrays, these quantities are usually defined in terms of the array factor f. The power radiated at infinity is, however, better modeled by the far field pattern $\boldsymbol{E}_\infty$ which differs from the array factor by the term $\hat{\boldsymbol{x}} \times (\hat{\boldsymbol{p}} \times \hat{\boldsymbol{x}})$. Instead of using $f = f(\hat{\boldsymbol{x}})$ in the following definitions one can equally well take $\alpha(\hat{\boldsymbol{x}})\, f(\hat{\boldsymbol{x}})$ where $\alpha(\hat{\boldsymbol{x}}) = \big|\hat{\boldsymbol{x}} \times (\hat{\boldsymbol{p}} \times \hat{\boldsymbol{x}})\big|$. We note again that $\alpha(\hat{\boldsymbol{x}}) = \sin\theta$ in polar coordinates if $\hat{\boldsymbol{p}} = \hat{\boldsymbol{e}}_3$. To follow the standard notations used in antenna theory, however, we take the array factor f for the definitions of these quantities.

We begin with the notion of directivity.

Definition 1.1. *Let $f = f(\hat{\boldsymbol{x}})$, $\hat{\boldsymbol{x}} \in S^2$, be the factor of an antenna array. We define the* **directivity** D_f *by*

$$D_f(\hat{\boldsymbol{x}}) \;:=\; \frac{|f(\hat{\boldsymbol{x}})|^2}{\frac{1}{4\pi}\int_{S^2} |f(\hat{\boldsymbol{x}}')|^2\, dS}\,, \quad \hat{\boldsymbol{x}} \in S^2\,, \tag{1.20}$$

if $f \neq 0$.

We write also $D_f(\theta,\phi)$ using spherical coordinates and suppress ϕ if f, and therefore D_f, is independent of ϕ.

This quantity D_f is sometimes called the **geometric** directivity (see [27]) since it is a quantity which depends only on the geometrical parameters of the antenna and not on the feeding mechanism.

The definition of directivity is a theoretical quantity and does not take into account the losses of power due to feeding mechanisms. In other words, our definition of directivity ignores the question of antenna input impedance due to the coupling of the power source with the antenna through a transmission line or wave guide.

In the case of a linear array along the x_3−axis, the array factor f is independent of ϕ and (1.20) reduces to

$$D_f(\theta) \;=\; \frac{|f(\theta)|^2}{\frac{1}{2}\int_0^\pi |f(\theta')|^2 \sin\theta'\, d\theta'}\,, \qquad 0 \leq \theta \leq \pi\,. \tag{1.21}$$

If we want to express explicitly the dependence of D_f on the feeding coefficients a_n we write $D_{\boldsymbol{a}}$ and use the operator $\mathcal{K} : \mathbb{C}^{2N+1} \longrightarrow C(S^2)$ and the vector notation again, see (1.19). It follows that we can express $|f(\hat{\boldsymbol{x}})|^2$ as a quadratic form

$$|f(\hat{\boldsymbol{x}})|^2 \;=\; |(\mathcal{K}\boldsymbol{a})(\hat{\boldsymbol{x}})|^2 \;=\; |(\boldsymbol{a}, \boldsymbol{e}(\hat{\boldsymbol{x}}))|^2 \;=\; (\boldsymbol{a}, \boldsymbol{C}(\hat{\boldsymbol{x}})\,\boldsymbol{a}) \tag{1.22}$$

where $\boldsymbol{C}(\hat{\boldsymbol{x}})$ is the Hermitian, positive semi-definite, $(2N+1) \times (2N+1)$ matrix with elements

$$c_{p,q}(\hat{\boldsymbol{x}}) \;=\; \boldsymbol{e}_p(\hat{\boldsymbol{x}})\,\overline{\boldsymbol{e}_q(\hat{\boldsymbol{x}})} \;=\; \mathrm{e}^{ik(\boldsymbol{y}_p - \boldsymbol{y}_q)\cdot(\hat{\boldsymbol{x}} - \hat{\boldsymbol{x}}_0)}\,. \tag{1.23}$$

Likewise, we introduce the matrix $\boldsymbol{B}$ with entries

$$b_{p,q} \;=\; \frac{1}{4\pi} \int\limits_{S^2} c_{p,q}\, dS \;=\; \frac{1}{4\pi} \int\limits_{S^2} \mathrm{e}^{ik(\boldsymbol{y}_p - \boldsymbol{y}_q)\cdot(\hat{\boldsymbol{x}} - \hat{\boldsymbol{x}}_0)}\, dS(\hat{\boldsymbol{x}})$$

which can be computed explicitly. Indeed, we make the change of variables $\hat{\boldsymbol{x}} = \boldsymbol{Q}^\top \hat{\boldsymbol{z}}$ where $\boldsymbol{Q} \in \mathbb{R}^{3\times 3}$ is the rotation which transforms $\boldsymbol{y}_p - \boldsymbol{y}_q$ into $|\boldsymbol{y}_p - \boldsymbol{y}_q|\,\hat{e}_3$ (the "north pole") and which yields

$$\begin{aligned}
b_{p,q} &= \frac{1}{4\pi}\, \mathrm{e}^{-ik(\boldsymbol{y}_p - \boldsymbol{y}_q)\cdot\hat{\boldsymbol{x}}_0} \int\limits_{S^2} \mathrm{e}^{ik(\boldsymbol{y}_p - \boldsymbol{y}_q)^\top \boldsymbol{Q}^\top \hat{\boldsymbol{z}}}\, dS \\
&= \frac{1}{4\pi}\, \mathrm{e}^{-ik(\boldsymbol{y}_p - \boldsymbol{y}_q)\cdot\hat{\boldsymbol{x}}_0} \int\limits_{S^2} \mathrm{e}^{-ik|\boldsymbol{y}_p - \boldsymbol{y}_q|\,\hat{e}_3\cdot\hat{\boldsymbol{z}}}\, dS \\
&= \frac{1}{2}\, \mathrm{e}^{-ik(\boldsymbol{y}_p - \boldsymbol{y}_q)\cdot\hat{\boldsymbol{x}}_0} \int\limits_0^\pi \mathrm{e}^{-ik|\boldsymbol{y}_p - \boldsymbol{y}_q|\cos\theta}\, \sin\theta\, d\theta \\
&= \frac{1}{2}\, \mathrm{e}^{-ik(\boldsymbol{y}_p - \boldsymbol{y}_q)\cdot\hat{\boldsymbol{x}}_0} \int\limits_{-1}^1 \mathrm{e}^{-ik|\boldsymbol{y}_p - \boldsymbol{y}_q|t}\, dt \\
&= \mathrm{e}^{-ik(\boldsymbol{y}_p - \boldsymbol{y}_q)\cdot\hat{\boldsymbol{x}}_0} \begin{cases} \dfrac{\sin\big(k\,|\boldsymbol{y}_p - \boldsymbol{y}_q|\big)}{k\,|\boldsymbol{y}_p - \boldsymbol{y}_q|}\,, & p \neq q, \\ \quad 1, & p = q. \end{cases}
\end{aligned} \tag{1.24}$$

The denominator of the expression (1.20) can then be written as an Hermitian form $(\boldsymbol{a}, \boldsymbol{B}\boldsymbol{a})$ so that the geometric directivity for an array takes the simple form

$$D_f(\hat{\boldsymbol{x}}) \;=\; D_{\boldsymbol{a}}(\hat{\boldsymbol{x}}) \;=\; \frac{(\boldsymbol{a}, \boldsymbol{C}(\hat{\boldsymbol{x}})\,\boldsymbol{a})}{(\boldsymbol{a}, \boldsymbol{B}\boldsymbol{a})}\,. \tag{1.25}$$

We note that $\boldsymbol{B}$ is positive definite and that both the matrices $\boldsymbol{C}(\hat{\boldsymbol{x}})$ and $\boldsymbol{B}$ depend on the parameters k, $\hat{\boldsymbol{x}}_0$ and $\boldsymbol{y}_n$.

For the linear equi-spaced array with spacing d, we have $\boldsymbol{y}_p - \boldsymbol{y}_q = d(p-q)\hat{\boldsymbol{e}}_3$ and thus

$$c_{p,q}(\theta) = \mathrm{e}^{ikd(p-q)(\cos\theta - \cos\theta_0)} \quad \text{and}$$

$$b_{p,q} = \mathrm{e}^{ikd(q-p)\cos\theta_0} \begin{cases} \dfrac{\sin(kd(p-q))}{kd(p-q)}\,, & p \neq q, \\ 1, & p = q\,. \end{cases}$$

In this case, both of the matrices $\boldsymbol{B}$ and $\boldsymbol{C}(\theta)$ are circulant, i.e. the entries $b_{p,q}$ and $c_{p,q}(\theta)$ depend only on $p-q$.

For circular arrays with radius a we have

$$\left|\boldsymbol{y}_p - \boldsymbol{y}_q\right| \;=\; a\sqrt{2(1-\cos(\phi_p - \phi_q)} \;=\; 2a\,\sin\frac{|p-q|\,\pi}{N}\,, \quad p, q = 1, \dots, N\,,$$

and thus

$$c_{p,q}(\hat{\boldsymbol{x}}) \;=\; \mathrm{e}^{ik(\boldsymbol{y}_p - \boldsymbol{y}_q)\cdot \boldsymbol{z}}$$

where $\boldsymbol{z}$ is the projection of $\hat{\boldsymbol{x}} - \hat{\boldsymbol{x}}_0$ onto the plane of the array, see (1.12). With plane polar coordinates $\boldsymbol{z} = \rho(\cos\xi, \sin\xi, 0)^\top$ this yields

$$c_{p,q}(\hat{\boldsymbol{x}}) \;=\; \mathrm{e}^{ik\rho\left[\cos(\phi_p - \xi) - \cos(\phi_q - \xi)\right]}$$

and

$$b_{p,q} \;=\; \mathrm{e}^{ik\sin\theta_0\left[\cos(\phi_q - \phi_0) - \cos(\phi_p - \phi_0)\right]}\,. \begin{cases} \dfrac{\sin\left(2ka\sin\frac{|p-q|\pi}{N}\right)}{2ka\sin\frac{|p-q|\pi}{N}}\,, & p \neq q, \\ 1\,, & p = q\,. \end{cases}$$

As an example, we compute the directivity for linear broadside arrays. In this way we will have another comparison of the effect of suppressing the side lobe level on the main beam.

Example 1.2. We consider a linear broadside array in the broadside direction (i.e. $\theta = \theta_0 = \pi/2$). We assume the inter-element spacing to be $d = \lambda/2$, i.e. $kd = \pi$. We denote the directivity for uniform, triangular, and binomial feeding by $D^U_{\boldsymbol{a}}$, $D^T_{\boldsymbol{a}}$ and $D^B_{\boldsymbol{a}}$, respectively. Then the matrices $\boldsymbol{C}(\pi/2)$ and $\boldsymbol{B}$ have a particularly simple form. Indeed, $\boldsymbol{B} = \boldsymbol{I}$ and $\boldsymbol{C}(\pi/2)$ is a full matrix whose entries are all 1. Naturally, the simplest case is that of the uniformly fed broadside array (see (1.7) for $\theta_0 = \pi/2$). In this case, $\boldsymbol{a} = \frac{1}{2N+1}(1, \dots, 1)^\top$ and thus

$$D^U_{\boldsymbol{a}}(\pi/2) \;=\; \frac{(2N+1)^2}{2N+1} \;=\; 2N+1. \tag{1.26}$$

Similarly, we compute the directivities for the triangular and binomial feedings. For $N = 3$, i.e. a seven element array, we have for the uniform feeding $D^U_{\boldsymbol{a}}(\pi/2) = 7$, while those for the triangular and binomial feedings are $D^T_{\boldsymbol{a}}(\pi/2) = 5.8182$ and $D^B_{\boldsymbol{a}}(\pi/2) = 4.4329$, respectively. These results illustrate once again that, in general, the attempt to suppress side lobes is met with a degradation of the directivity of the array.

We will now introduce other measures of performance, the *signal-to-noise ratio*, denoted by *SNR*, and the *radiation efficiency*, which we will denote by G. The signal-to-noise ratio is defined in terms of the antenna factor alone:

Definition 1.3. *Let $f = f(\hat{\boldsymbol{x}}) \neq 0$ be the antenna factor and $\omega \in L^\infty(S^2)$ the noise temperature. Then we define the* **signal-to-noise ratio** *by*

$$\mathit{SNR}_f(\hat{\boldsymbol{x}}) \;:=\; \frac{|f(\hat{\boldsymbol{x}})|^2}{\frac{1}{4\pi}\int_{S^2} |f(\hat{\boldsymbol{x}}')|^2\, \omega(\hat{\boldsymbol{x}}')^2\, dS}\,, \quad \hat{\boldsymbol{x}} \in S^2\,, \tag{1.27}$$

if $f \neq 0$.

The denominator represents relative noise power. In terms of the feeding operator $\mathcal{K} : \mathbb{C}^{2N+1} \longrightarrow C(S^2)$ and in vector notation, the signal-to-noise ratio takes the form

$$\mathit{SNR}_{\boldsymbol{a}}(\hat{\boldsymbol{x}}) \;:=\; \frac{|(\mathcal{K}\boldsymbol{a})(\hat{\boldsymbol{x}})|^2}{\frac{1}{4\pi}\int_{S^2} |(\mathcal{K}\boldsymbol{a})(\hat{\boldsymbol{x}}')|^2\, \omega(\hat{\boldsymbol{x}}')^2\, dS} \;=\; \frac{(\boldsymbol{a}, \boldsymbol{C}(\hat{\boldsymbol{x}})\, \boldsymbol{a})}{(\boldsymbol{a}, \boldsymbol{N}\boldsymbol{a})}\,, \quad \hat{\boldsymbol{x}} \in S^2\,, \tag{1.28}$$

if $\boldsymbol{a} \neq 0$. Note that the matrix C has the form (1.23). The elements of the (positive definite) noise matrix $\boldsymbol{N}$, which we write as $n_{p,q}$, are given by

$$n_{p,q} \;:=\; \frac{1}{4\pi} \int\limits_{S^2} e^{ik(\boldsymbol{y}_p - \boldsymbol{y}_q)\cdot(\hat{\boldsymbol{x}} - \hat{\boldsymbol{x}}_0)}\, \omega(\hat{\boldsymbol{x}})^2\, dS(\hat{\boldsymbol{x}})\,. \tag{1.29}$$

We will give a more detailed discussion of the SNR-functional in Chapter 7.

In contrast to the directivity and signal-to-noise ratio, the **efficiency index** $G_{\boldsymbol{a}}$ depends explicitly on the feeding coefficients $\boldsymbol{a}$. It is defined by

$$G_{\boldsymbol{a}}(\hat{\boldsymbol{x}}) \;:=\; \frac{|f(\hat{\boldsymbol{x}})|^2}{|\boldsymbol{a}|^2} \;=\; \frac{(\boldsymbol{a}, \boldsymbol{C}(\hat{\boldsymbol{x}})\, \boldsymbol{a})}{|\boldsymbol{a}|^2}\,, \quad \hat{\boldsymbol{x}} \in S^2\,, \tag{1.30}$$

if $\boldsymbol{a} \neq 0$.

It is common to refer to the ratio of the directivity to the radiation efficiency as the quality factor of an array.

Definition 1.4. *Let $f = f(\hat{x}) \neq 0$ be the antenna factor of an array with feeding coefficient $\boldsymbol{a} = (a_{-N}, \dots, a_N)^\top \in \mathbb{C}^{2N+1}$. Then we define the* **quality factor** *(or Q-factor) by*

$$Q_{\boldsymbol{a}} := \frac{|\boldsymbol{a}|^2}{\|f\|^2_{L^2(S^2)}} = \frac{|\boldsymbol{a}|^2}{(\boldsymbol{a}, \boldsymbol{B}\boldsymbol{a})} . \tag{1.31}$$

Note that the matrix $\boldsymbol{B}$ has the form (1.24).

Intuitively, the Q-factor measures the proportion of input power which fails to be radiated into the far field. As such, it would be advantageous to make this factor as small as possible. In the next subsection we will see, however, that in general, an increase in directivity is accompanied by a corresponding increase in the Q-factor so that the antenna fails to radiate power efficiently.

1.3.2 Maximization of Directivity

For the case of a finite array we have expressed the directivity $D_{\boldsymbol{a}}$, the signal-to-noise-ratio *SNR*, and the Q-factor by ratios of quadratic forms (see (1.25), (1.28), and (1.31), respectively). For the optimization of these we recall the following result from linear algebra.

Theorem 1.5. *Let $\boldsymbol{C}, \boldsymbol{B} \in \mathbb{C}^{n\times n}$ be Hermitian and positive semi-definite matrices with $\boldsymbol{B}$ positive definite. Let $R(\boldsymbol{a}) = \frac{(\boldsymbol{a},\boldsymbol{C}\boldsymbol{a})}{(\boldsymbol{a},\boldsymbol{B}\boldsymbol{a})}$ for $\boldsymbol{a} \neq 0$. Then the maximum value for R is the largest eigenvalue μ of the* **generalized eigenvalue problem***:*

$$\boldsymbol{C}\boldsymbol{v} = \mu \boldsymbol{B}\boldsymbol{v} . \tag{1.32}$$

We should mention that some authors suggest that, since the matrix $\boldsymbol{B}$ is positive definite and hence invertible, the optimal quantities can be expressed in terms of the usual eigenvalue problem for the matrix $\boldsymbol{B}^{-1}\boldsymbol{C}$. However, it is well known (see [144]), that computation directly with the generalized eigenvalue problem using, for example, the QZ algorithm is in general more stable and leads to more accurate results.

Example 1.6. As we mentioned above, in the case of the broadside array with element spacing $d = \lambda/2$ the matrix $\boldsymbol{B}$ has a particularly simple form, namely $\boldsymbol{B} = \boldsymbol{I}$ and the matrices $\boldsymbol{C}(\theta)$ are circulant. It is easy to see that in this case there exists only one non-zero eigenvalue, namely $\mu = 2N + 1$, and that the corresponding eigenvector $\boldsymbol{v}$ is given by $v_q = \exp(ikdq \cos\theta)$, $q = -N, \dots, N$. Therefore, the uniform feeding is only optimal for $\theta = \pi/2$, i.e. the broadside direction, but the optimal value is always $2N + 1$. For other spacings, the $\boldsymbol{B}$ matrix is more complicated. We have made the computation for three and seven element broadside and end-fire arrays for spacings from $d/\lambda = 0.1$ to $d/\lambda = 1$. We present the maximal values

$$D_{max}(\theta) := \max\{D_{\boldsymbol{a}}(\theta) : \boldsymbol{a} \in \mathbb{C}^{2N+1},\ \boldsymbol{a} \neq 0\}$$

of the directivities in the tables below (Figures 1.10 and 1.11) together with the corresponding Q-factors as a function of d/λ. Note the dramatic increase in the size of Q_a as the spacing tends to zero. Thus, the fraction of power fed to the antenna which is radiated into the far field becomes very small and the antenna is very inefficient for small values of d.

	3 elements		7 elements	
d/λ	$D_{max}(\pi/2)$	Q	$D_{max}(\pi/2)$	Q
0.1	2.2714	219.3716	4.8489	1.5×10^8
0.2	2.3394	10.7757	5.0500	2.3×10^4
0.3	2.4657	1.8732	5.4211	74.8379
0.4	2.6737	0.9819	6.0301	1.4250
0.5	3.0000	1.0000	7.0000	1.0000
0.6	3.4800	1.1731	8.3312	1.1922
0.7	4.0396	1.3605	9.5777	1.3786
0.8	4.2513	1.4206	10.6327	1.5369
0.9	3.7255	1.2419	11.4244	1.6436
1.0	3.0000	1.0000	7.0000	1.0000

Fig. 1.10. Optimal Values $D_{max}(\pi/2)$ and Corresponding Q for 3– and 7–Element Broadside Arrays

	3 elements		7 elements	
d/λ	$D_{max}(0)$	Q	$D_{max}(0)$	Q
0.1	8.7283	244.9548	47.4029	1.66×10^8
0.2	7.9034	16.4239	42.4906	3.49×10^7
0.3	6.5173	3.8458	33.8826	2.19×10^2
0.4	4.6823	1.6626	21.0384	6.6948
0.5	3.0000	1.0000	7.0000	1.0000
0.6	2.5824	0.8896	5.9381	0.8792
0.7	2.9562	1.0198	6.8266	1.0163
0.8	3.2798	1.1437	7.7255	1.1464
0.9	3.5014	1.1829	8.5107	1.2652
1.0	3.0000	1.0000	7.0000	1.0000

Fig. 1.11. Optimal Values $D_{max}(0)$ and Corresponding Q for 3– and 7–Element Broadside Arrays

The problem of avoiding large Q-factors leads naturally to a problem of constrained optimization in which we can ask for the current inputs which will maximize the directivity subject to the constraint that the Q-factor is kept at or below a preassigned value. Other constraints may be imposed as well. We show here how linearly constrained optimization problems are related to the generalized eigenvalue problem.

The general problem of maximizing the ratio of quadratic forms is

$$\begin{aligned} &\text{Maximize} \quad \frac{(\boldsymbol{a}, \boldsymbol{C}\boldsymbol{a})}{(\boldsymbol{a}, \boldsymbol{B}\boldsymbol{a})}, \\ &\text{subject to} \quad (\boldsymbol{z}_j, \boldsymbol{a}) = 0, \quad j = 1, \dots, m. \end{aligned} \tag{1.33}$$

Here $\boldsymbol{C}$ and $\boldsymbol{B}$ are Hermitian positive semi-definite $n \times n$–matrices, $\boldsymbol{B}$ positive definite. (In our application $n = 2N + 1$.) Suppose that $Z := \text{span}\{\boldsymbol{z}_1, \boldsymbol{z}_2, \dots, \boldsymbol{z}_m\}$, and that $\mathbb{C}^n$ is decomposed into the orthogonal sum Z and Y where a basis for Y is given by $\{\boldsymbol{y}_1, \boldsymbol{y}_2, \dots, \boldsymbol{y}_{n-m}\}$. Since $\boldsymbol{a}$ is constrained to be orthogonal to the subspace Z it has to be in Y so that the vector $\boldsymbol{a} \in Y$ can be expressed as $\boldsymbol{a} = \boldsymbol{W}\boldsymbol{c}$ where $\boldsymbol{W}$ is an $n \times (n-m)$ matrix whose columns are the $\boldsymbol{y}_i$ and $\boldsymbol{c}$ is an $(n-m)$-vector. Hence the form (1.33) becomes

$$\text{Maximize} \quad \frac{(\boldsymbol{c}, \boldsymbol{W}^* \boldsymbol{C}\, \boldsymbol{W} \boldsymbol{c})}{(\boldsymbol{c}, \boldsymbol{W}^* \boldsymbol{B}\, \boldsymbol{W} \boldsymbol{c})}. \tag{1.34}$$

As a practical matter for finding a basis for the subspace Y, we can apply a **Householder transformation** to the matrix $\boldsymbol{U}$ whose columns are formed by the vectors $\boldsymbol{z}_j$. If $\boldsymbol{H}$ is a Householder matrix which puts $\boldsymbol{U}$ in Hessenberg form

$$\begin{pmatrix} \boldsymbol{R} \\ \boldsymbol{O} \end{pmatrix}$$

where $\boldsymbol{R}$ is an $m \times m$ matrix which is tridiagonal, then the last $n - m$ rows of $\boldsymbol{H}$ are linearly independent and form a basis for the subspace Y [144]. We may take $\boldsymbol{W}$ in (1.34) to be the $n \times (n - m)$ matrix with these rows. With this choice, we can easily check that the quadratic forms are non-negative and that the positive definiteness of $\boldsymbol{B}$ implies that of $\boldsymbol{W}^* \boldsymbol{B}\, \boldsymbol{W}$. We conclude that the linearly constrained problem reduces to a generalized eigenvalue problem of the same type as discussed above.

1.4 Dolph-Tschebysheff Arrays

We have seen in previous sections that, for a linear array of dipoles, it is possible to affect the side lobe level in a variety of ways by means of choosing various inputs to the sources. Indeed, we have seen in Subsection 1.2.1 that, with a binomial distribution, we are able to suppress the side lobes entirely. However, we have also seen that lowering or even eliminating the side lobe power comes at expense of increasing the beam-width and reducing the directivity of the main lobe. At the risk of repeating ourselves, we see in this situation that there is a trade-off between beam-width and side lobe level.

This fact led Dolph [34] to pose and solve the problem of finding the current distribution which yields the narrowest main beam-width under the constraint that the side lobes do not exceed a fixed level. In this section we will present this optimization problem for broadside arrays. Dolph's solution depends on certain properties of the Tschebysheff polynomials which we present in the next part of this section.

1.4.1 Tschebysheff Polynomials

There are many equivalent ways to define the Tschebysheff polynomials of the first kind. On the interval $[-1,1]$ the **Tschebysheff polynomial** T_n of order n can be defined explicitly by the relation:

$$T_n(x) \;=\; \cos(n \arccos x), \quad -1 \le x \le 1, \quad n = 1, 2, \ldots \tag{1.35}$$

This definition shows immediately that

$$\max_{-1\le x\le 1} |T_n(x)| \;=\; T_n(1) \;=\; (-1)^n\, T_n(-1) \;=\; 1. \tag{1.36}$$

Moreover, the cosine addition formula shows that the polynomials obey the recursion formula

$$T_{n+1}(x) \;+\; T_{n-1}(x) \;=\; 2x\, T_n(x)\,. \tag{1.37}$$

From this recursion relation it is easy to see that, since $T_0(x) = 1$ and $T_1(x) = x$, the T_n are polynomials in x of degree n with leading coefficient 2^{n-1} for $n \ge 1$, and hence can be extended to the entire real line. Likewise, from the recursion relation it is evident that the polynomials of odd order contain only odd powers of the variable x while the polynomials of even order contain only even powers of that variable. The substitution $x := \cos\theta$ then results in the relation

$$\int\limits_{-1}^{1} T_n(x)\, T_m(x)\, \frac{dx}{\sqrt{1-x^2}} \;=\; \int\limits_{0}^{\pi} \cos(n\theta)\, \cos(m\theta)\, d\theta \;=\; \begin{cases} \pi, & n = m = 0, \\ \pi/2, & n = m \neq 0, \\ 0, & n \neq m, \end{cases} \tag{1.38}$$

and so these polynomials form an orthogonal system with respect to the weight $1/\sqrt{1-x^2}$. The system $\{T_n : n = 0, 1, \ldots\}$ is complete in the Hilbert space $L^2(-1,1)$ as well as in the space $C[-1,1]$. Figure 1.12 shows the graphs of $T_n(x)$ for $n = 1, 2, 3, 4$.

The graphs of the Tschebysheff polynomial suggest certain important facts. Looking at the form (1.35) the zeros of these polynomials are given by the roots of the equation

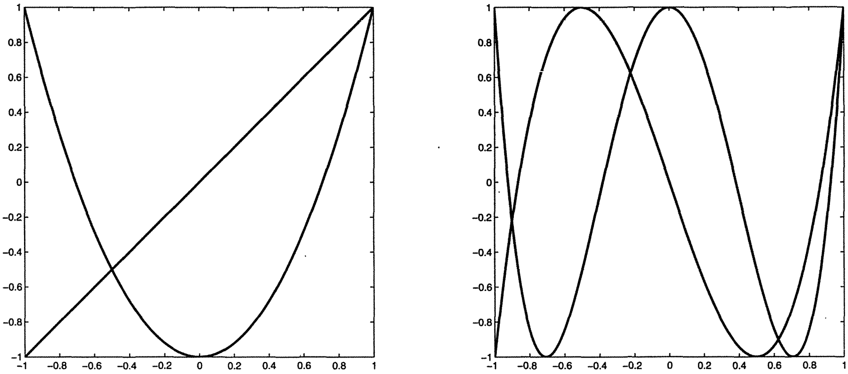


Fig. 1.12. Graph of Tschebysheff Polynomials T_1, T_2 (left) and T_3, T_4 (right)

$$\cos(n\theta) \;=\; 0,$$

or by

$$x_k \;=\; \cos\left(\frac{(2k-1)\pi}{2n}\right), \quad k = 1, 2, \ldots, n. \tag{1.39}$$

It is also easy to compute the critical points of T_n, those points being solutions of the equation $\sin(n\theta) = 0$. Equivalently, the critical points are

$$\tilde{x}_k \;=\; \cos(k\pi/n), \quad k = 1, 2, \ldots, n, \tag{1.40}$$

and at these latter points,

$$T_n(\tilde{x}_k) \;=\; T_n\big(\cos(k\pi/n)\big) \;=\; \cos(nk\pi/n) \;=\; (-1)^k. \tag{1.41}$$

We should keep in mind that both the set of critical points and the set of zeros are subsets of the interval $[-1,1]$. Moreover it should also be clear that the extended functions T_n are monotonic outside of the interval $[-1,1]$. Specifically, for $x > 1$, T_n is monotonically increasing, while for $x < -1$, T_n is monotonically decreasing or increasing depending upon whether n is even or odd.

The property of the Tschebysheff polynomials which is of most interest to us here is the following remarkable optimality property in the space $C[-1,1]$, equipped with the norm of uniform convergence, and which was discovered by Tschebysheff.

Theorem 1.7. *Of all polynomials of degree n with leading coefficient* 1, *the polynomial with the smallest maximum norm on* $[-1,1]$ *is* $2^{1-n}T_n$.

There are several similar theorems see, e.g.,[2], Chapter II. We will need the following version for *even* polynomials. Recalling that the largest zero of T_n is $\hat{x} = \cos\frac{\pi}{2n}$ we can state it as:

Theorem 1.8. *Let n be an even integer, $x_0 \in [-1,0]$ and $\beta > \hat{x}$, the largest zero of the polynomial T_n. Then*

$$p^* := \frac{1}{T_n(\beta)}\, T_n \tag{1.42}$$

is the unique solution of the minimization problem

$$(\mathcal{P}_{DT}) \qquad \begin{array}{l} \textit{Minimize} \quad \max\limits_{x_0 \le x \le \hat{x}} |p(x)| \\ \textit{Subject to}: p \in \mathcal{P}_n,\ p \text{ even},\ p(\hat{x}) = 0,\ p(\beta) = 1. \end{array}$$

Here, $\mathcal{P}_n$ denotes the space of algebraic polynomials of order at most n.

Proof: Assume, on the contrary, that there exists some admissible polynomial p such that

$$\max_{x_0 \le x \le \hat{x}} |p(x)| \;<\; \max_{x_0 \le x \le \hat{x}} |p^*(x)| \;=\; \frac{1}{T_n(\beta)}\,.$$

Then the polynomial $q := p^* - p$ vanishes at both $\hat{x}$ and β. Since the maximum of $|p|$ is smaller than the maximum of $|p^*|$, the polynomial q has alternating signs at the successive extreme points of T_n, namely at $\cos\frac{k\pi}{n}$, $k = 1, 2, \ldots, \frac{n}{2}$. Thus there are $\frac{n}{2}-1$ zeros of q in the interval $(0, \hat{x})$. Considering that q vanishes at both $\hat{x}$ and β, it has $\frac{n}{2}+1$ positive zeros and, since it is even, $n+2$ real zeros. Since the polynomial q has degree at most n, this contradicts the fact that q can have at most n roots[2] and the proof is complete. □

We remark that, in the formulation of the optimization problem, we do not require that the admissible polynomials have no zeros larger than $\hat{x}$. However, it turns out "automatically" that the optimal polynomial p^* enjoys this property.

1.4.2 The Dolph Problem

The optimization problem considered by Dolph in his famous paper [34] can be stated in the following terms.

For a given side lobe level and beam-width of the main lobe i.e., twice the distance (measured in degrees) from the center of the beam to the first null, maximize the peak power in the main lobe.

An equivalent statement is the following:

For a given main beam-width and peak power in the main beam, minimize the level of the side lobes in the sense of minimizing the maximum value of the array pattern i.e., the magnitude of the array factor outside the main beam.

Indeed, these statements are equivalent in the sense that the same excitations lead to the optimal solutions. We take the second formulation in order to make

[2] The fact that a polynomial of degree n can have at most n roots is known as the Fundamental Theorem of Algebra

a precise statement of the problem. In particular, we will restrict ourselves to the standard example of cophasal, symmetric excitations in which case the coefficients a_n can be taken as real numbers with $a_{-n} = a_n$ and the expression for the array factor takes on the form (1.11). Moreover, since the array factor is symmetric with respect to $\theta = \pi/2$ (i.e., $f(\pi/2 - \theta) = f(\pi/2 + \theta)$ for $0 \le \theta \le \pi$) we may restrict our consideration to the interval $[0, \pi/2]$.

If we introduce the set $\mathcal{T}_N$ of even trigonometric polynomials in the variable $\gamma = kd\cos\theta$, i.e.

$$\mathcal{T}_N \;=\; \left\{ a_0 + 2\sum_{n=1}^{N} a_n \cos(n\gamma) \,:\, a_n \in \mathbb{R} \right\}, \tag{1.43}$$

we can then state the optimization problem for the linear array as the following:

Let $\hat{\theta} \in (0, \pi/2)$ be fixed.

$$(\mathcal{P}_{DT}) \qquad \begin{cases} \text{Minimize} \quad I(f) \;:=\; \max\limits_{0 \le \theta \le \hat{\theta}} |f(\theta)| \\ \text{Subject to} : f \in \mathcal{T}_N, \text{ and } f(\hat{\theta}) = 0,\; f(\pi/2) = 1. \end{cases}$$

In this optimization problem, we fix the maximum amplitude i.e., the peak power of the main beam to be 1 at $\pi/2$ and we fix $\hat{\theta}$ to be a null. We do not require that it is the largest null in $[0, \pi/2]$. However, it will turn out to be the largest null of the optimal solution f^o (see (1.47)).

Returning to the expression for the array pattern, recall that it has the form

$$f(\theta) \;=\; a_0 \;+\; 2\sum_{n=1}^{N} a_n \,\cos\bigl(n\gamma(\theta)\bigr), \tag{1.44}$$

where $\gamma(\theta) = kd\cos\theta$. From $\cos\gamma = 2\cos^2(\gamma/2) - 1$ and $\cos(n\gamma) = T_n(\cos\gamma) = T_n\bigl(2\cos^2(\gamma/2) - 1\bigr)$ it follows that the array pattern (1.44) can be written as

$$f(\theta) \;=\; a_0 \;+\; 2\sum_{n=1}^{N} a_n\, T_n\bigl(2\cos^2(\gamma/2) - 1\bigr)$$

which is an algebraic polynomial in $\cos^2(\gamma/2)$. Therefore, if $\mathcal{P}_{2N}$ denotes the real vector space of algebraic polynomials of degree at most $2N$, we have

$$\mathcal{T}_N \;=\; \bigl\{ p\bigl(\cos(\gamma/2)\bigr) : p \in \mathcal{P}_{2N},\; p \text{ even} \bigr\}. \tag{1.45}$$

We now transform the problem $(\mathcal{P}_{DT})$ into a minimization problem over the set of even algebraic polynomials by setting $x = \beta\cos\bigl(\frac{kd}{2}\cos\theta\bigr)$. We choose the parameter β so that $\theta = \hat{\theta}$ is mapped to $x = \hat{x} = \cos\frac{\pi}{4N}$ which is the largest zero of the Tschebysheff polynomial T_{2N}. This requires β to be

$$\beta = \frac{\cos\frac{\pi}{4N}}{\cos(\frac{kd}{2}\cos\hat{\theta})}. \qquad (1.46)$$

Note that, under this transformation, the points $\theta = 0$ and $\theta = \pi/2$ are mapped into $x = \beta\cos\frac{kd}{2}$ and $x = \beta$ respectively.

Now setting $x_0 = \beta\cos\frac{kd}{2}$ we can rewrite the problem $(\mathcal{P}_{DT})$ in the form

$$\begin{aligned} &\text{Minimize:} \quad \max_{x_0 \le x \le \hat{x}} |p(x)| \\ &\text{Subject to:} \quad p \in \mathcal{P}_{2N},\ p \text{ even},\ p(\hat{x}) = 0,\ p(\beta) = 1. \end{aligned}$$

We can now apply Theorem 1.8 provided that: (i) $\beta > \hat{x}$ and (ii) $-1 \le x_0 \le 0$. The first of these conditions is guaranteed provided $0 < \cos(\frac{kd}{2}\cos\hat{\theta}) < 1$ i.e., provided $0 < \frac{kd}{2}\cos\hat{\theta} < \pi/2$. This is equivalent to the requirement that

$$d\cos\hat{\theta} < \frac{\lambda}{2}.$$

The second requirement holds provided $\cos\frac{kd}{2} \le 0$ and $\beta\cos\frac{kd}{2} \ge -1$. The first of these two inequalities holds provided $\pi/2 \le \frac{kd}{2} \le 3\pi/2$, which is equivalent to

$$\frac{\lambda}{2} \le d \le \frac{3\lambda}{2}.$$

Under these conditions, the application of Theorem 1.8 leads to the result that the optimal solution of the problem $(\mathcal{P}_{DT})$ is

$$f^o(\theta) = \frac{1}{T_{2N}(\beta)}\, T_{2N}\left(\beta\cos\left(\frac{kd}{2}\cos\theta\right)\right) \qquad (1.47)$$

with optimal value equal to $1/T_{2N}(\beta)$. Here, β is given by (1.46) This optimal value is certainly less than 1 if $\beta > 1$. In the case that $\beta < 1$ the form f^* also gives the optimal solution but in this case, the side lobe level is larger than 1 (since in this case $T_{2N}(\beta) < 1$). This represents a non-physical case and is avoided if N is large or the spacing d is chosen so that $d\cos\hat{\theta} \simeq \lambda/2$.

In Figure 1.13 we show the pattern for $N = 3$, $d = \lambda/2$ and $\hat{\theta} = \frac{4}{5}\cdot\frac{\pi}{2}$, i.e. beam-width $\pi/5$.

1.5 Line Sources

The electromagnetic fields of a finite line current flowing along the curve $C \subset \mathbb{R}^3$ with parametrization $\boldsymbol{y} = \boldsymbol{y}(s)$, $a \le s \le b$, can be modeled as the limiting form of an array where the distance between the elements tends to

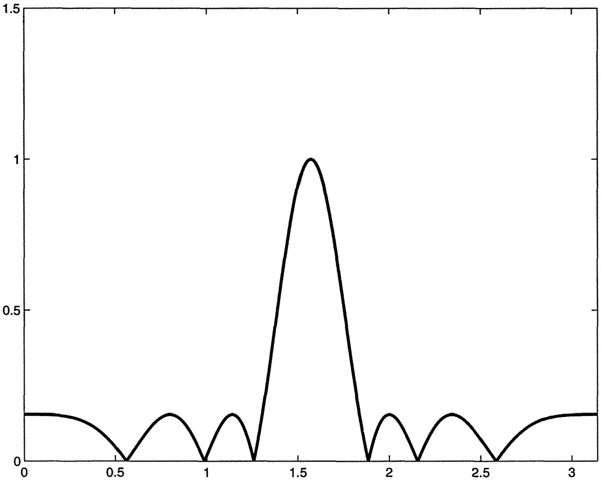


Fig. 1.13. Array for Optimal Solution of Dolph Problem ($N = 3$, $d = \lambda/2$ and $\hat{\theta} = \frac{4}{5} \cdot \frac{\pi}{2}$)

zero and the number of elements increases to infinity. Assuming the common dipole moment $\hat{\boldsymbol{e}}_3$ we derive an expression for the electric far field

$$\boldsymbol{E}_\infty(\hat{\boldsymbol{x}}) = \frac{i\omega\mu_0}{4\pi}\,[\hat{\boldsymbol{x}} \times (\hat{\boldsymbol{e}}_3 \times \hat{\boldsymbol{x}})] \int\limits_C \psi(\boldsymbol{y})\,\mathrm{e}^{-ik\boldsymbol{y}\cdot\hat{\boldsymbol{x}}}\,d\ell(\boldsymbol{y}) \tag{1.48a}$$

$$= \frac{i\omega\mu_0}{4\pi}\,[\hat{\boldsymbol{x}} \times (\hat{\boldsymbol{e}}_3 \times \hat{\boldsymbol{x}})] \int\limits_a^b \psi\big(\boldsymbol{y}(s)\big)\,\mathrm{e}^{-ik\boldsymbol{y}(s)\cdot\hat{\boldsymbol{x}}}\,\big|\dot{\boldsymbol{y}}(s)\big|\,ds\,, \tag{1.48b}$$

$\hat{\boldsymbol{x}} \in S^2$, and thus, using spherical polar coordinates,

$$|\boldsymbol{E}_\infty(\hat{\boldsymbol{x}})| \;=\; \frac{\omega\mu_0}{4\pi}\,\sin\theta\left|\int\limits_a^b \psi\big(\boldsymbol{y}(s)\big)\,\mathrm{e}^{-ik\boldsymbol{y}(s)\cdot\hat{\boldsymbol{x}}}\,\big|\dot{\boldsymbol{y}}(s)\big|\,ds\right|, \quad \hat{\boldsymbol{x}} \in S^2\,. \tag{1.49}$$

Analogous to the case of an array, we define

$$f(\hat{\boldsymbol{x}}) \;:=\; \int\limits_C \psi(\boldsymbol{y})\,\mathrm{e}^{-ik\boldsymbol{y}\cdot\hat{\boldsymbol{x}}}\,d\ell(\boldsymbol{y})\,, \quad \hat{\boldsymbol{x}} \in S^2\,, \tag{1.50}$$

and refer to the function f as the **line factor**.

Just as we did in the case of an array we can specify a particular direction $\hat{\boldsymbol{x}}_0 \in S^2$ and replace $\psi(\boldsymbol{y})$ by $\psi(\boldsymbol{y})\exp(ik\boldsymbol{y}\cdot\hat{\boldsymbol{x}}_0)$. This substitution yields

$$f(\hat{\boldsymbol{x}}) \;=\; \int\limits_C \psi(\boldsymbol{y})\,\mathrm{e}^{ik\boldsymbol{y}\cdot(\hat{\boldsymbol{x}}_0-\hat{\boldsymbol{x}})}\,d\ell(\boldsymbol{y})\,, \quad \hat{\boldsymbol{x}} \in S^2\,. \tag{1.51}$$

It is obvious that this concept of a line source is a continuous version of the array. Indeed, the replacement of the integral in (1.50) by a Riemann sum with quadrature points s_n, $n = -N, \dots, N$, reduces the line factor to the form (1.2) for an array with feeding coefficients $a_n = \psi(\boldsymbol{y}(s_n))\, |\dot{\boldsymbol{y}}(s_n)| \,/(2N+1)$. In spite of this close relationship between these two concepts there is an essential mathematical difference: While the feeding vectors $\boldsymbol{a}$ for arrays lie in the *finite dimensional* space $\mathbb{C}^{2N+1}$, the current ψ for the line source is an element of an *infinite dimensional* (function) *space.*

The definitions 1.20 and 1.27 of directivity and signal-to-noise ratio are independent of the feeding and make sense also for line sources. The feeding is now modeled by the operator which maps $\psi \in L^2(C)$ into the line factor f – or the far field pattern $\boldsymbol{E}_\infty$. In order to treat both cases simultaneously, we allow $\alpha \in C(S^2)$ to be an arbitrary function with $\alpha \geq 0$ on S^2 such that $\big\{\hat{\boldsymbol{x}} \in S^2 : \alpha(\hat{\boldsymbol{x}}) > 0\big\}$ is dense in S^2. We think of α being either constant 1 or $\alpha(\hat{\boldsymbol{x}}) = \big|\hat{\boldsymbol{x}} \times (\hat{\boldsymbol{e}}_3 \times \hat{\boldsymbol{x}})\big|$. Note that in the latter case α is independent of the angular variable ϕ and is given by $\alpha(\theta) = \sin\theta$ in polar coordinates.

We then define the corresponding far field operator $\mathcal{K} : L^2(C) \longrightarrow C(S^2)$ by

$$(\mathcal{K}\psi)(\hat{\boldsymbol{x}}) \;:=\; \alpha(\hat{\boldsymbol{x}}) \int\limits_C \psi(\boldsymbol{y})\, e^{ik\boldsymbol{y}\cdot(\hat{\boldsymbol{x}}_0 - \hat{\boldsymbol{x}})}\, d\ell(\boldsymbol{y})\,, \quad \hat{\boldsymbol{x}} \in S^2\,. \tag{1.52}$$

If the operator $\mathcal{K}$ is one-to-one[3] (or *injective*) we can formulate directivity and signal-to-noise ratio as:

$$D_\psi(\hat{\boldsymbol{x}}) \;=\; \frac{|(\mathcal{K}\psi)(\hat{\boldsymbol{x}})|^2}{\frac{1}{4\pi}\int_{S^2} |(\mathcal{K}\psi)(\hat{\boldsymbol{x}}')|^2\, dS} \;=\; \frac{|(\mathcal{K}\psi)(\hat{\boldsymbol{x}})|^2}{\frac{1}{4\pi}\,\|\mathcal{K}\psi\|^2_{L^2(S^2)}}\,, \quad \hat{\boldsymbol{x}} \in S^2\,, \tag{1.53a}$$

and

$$SNR_\psi(\hat{\boldsymbol{x}}) \;=\; \frac{|(\mathcal{K}\psi)(\hat{\boldsymbol{x}})|^2}{\frac{1}{4\pi}\int_{S^2} |(\mathcal{K}\psi)(\hat{\boldsymbol{x}}')|^2\, \omega(\hat{\boldsymbol{x}}')^2\, dS} \;=\; \frac{|(\mathcal{K}\psi)(\hat{\boldsymbol{x}})|^2}{\frac{1}{4\pi}\,\|\omega\,\mathcal{K}\psi\|^2_{L^2(S^2)}}\,, \tag{1.53b}$$

for $\hat{\boldsymbol{x}} \in S^2$ and $\psi \neq 0$.

The radiation efficiency and the quality factor are defined analogously as in the case of an array of point sources:

Definition 1.9. *Let the operator* $\mathcal{K} : L^2(C) \longrightarrow C(S^2)$ *be one-to-one. Then we define the* **efficiency index** G_ψ *and the* **quality factor** Q *by*

$$G_\psi(\hat{\boldsymbol{x}}) := \frac{|(\mathcal{K}\psi)(\hat{\boldsymbol{x}})|^2}{\|\psi\|^2_{L^2(C)}}\,, \quad \hat{\boldsymbol{x}} \in S^2\,, \quad \textit{and} \tag{1.54a}$$

$$Q_\psi := \frac{\|\psi\|^2_{L^2(C)}}{\|\mathcal{K}\psi\|^2_{L^2(S^2)}}\,. \tag{1.54b}$$

[3] We later give examples for the operator to be one-to-one (Theorem 1.15).

In the next subsections we will study the linear and the circular line sources in more detail.

We have seen that in the case of a finite array of point sources the directivity remains bounded if we vary the coefficients a_n. In the case of line sources this is not always the case due to the fact that the space of feeding coefficients is **infinite dimensional**. We note that we can write the directivity from (1.53a) in the form

$$D_\psi(\hat{\boldsymbol{x}}) = \frac{\left|(\psi,\rho)_{L^2(C)}\right|^2}{\frac{1}{4\pi}\,\|\mathcal{K}\psi\|^2_{L^2(S^2)}}\,,\quad \hat{\boldsymbol{x}} \in S^2\,,\quad \text{with } \rho(\boldsymbol{y}) = \alpha(\hat{\boldsymbol{x}})\,\mathrm{e}^{ik\boldsymbol{y}\cdot(\hat{\boldsymbol{x}}-\hat{\boldsymbol{x}}_0)}\,,\ \boldsymbol{y} \in C\,.$$

Then we can prove the following abstract theorem from functional analysis which describes the range $\mathcal{R}(L^*)$ of the adjoint L^* of an operator L:

Theorem 1.10. *Assume that $L : X \longrightarrow Y$ is linear, bounded, and one-to-one between Hilbert spaces X and Y, and $\rho \in X$. Then*

$$\rho \in \mathcal{R}(L^*) \qquad \textit{if and only if} \qquad \sup_{x\neq 0} \frac{|(\rho,x)|}{\|Lx\|} < \infty\,,$$

where again $L^* : Y \longrightarrow X$ denotes the adjoint of L, i.e. $(Lx,y)_Y = (x,L^*x)_X$ for all $x \in X$, $y \in Y$.

Proof: If $\rho = L^*z$ then, using the Cauchy-Schwarz inequality,

$$\frac{|(\rho,x)|}{\|Lx\|} = \frac{\left|(L^*z,x)\right|}{\|Lx\|} = \frac{|(z,Lx)|}{\|Lx\|} \leq \|z\| \tag{1.55}$$

for all $x \neq 0$.

To prove the converse statement, assume that $\rho \notin \mathcal{R}(L^*)$ and define the subspace $V := \big\{x \in X : (x,\rho) = 0\big\}$. First, we show that the orthogonal complements[4] of $L(V)$ and $L(X) = \mathcal{R}(L)$ coincide, i.e. $L(V)^\perp = L(X)^\perp$. The inclusion $L(V)^\perp \supset L(X)^\perp$ is obvious since $L(V) \subset L(X)$. Let $y \in L(V)^\perp$, i.e. $0 = (y,Lv) = (L^*y,v)$ for all $v \in V$. Thus $L^*y \in V^\perp = \operatorname{span}\{\rho\}$. The assumption $\rho \notin \mathcal{R}(L^*)$ yields $L^*y = 0$, i.e. $y \in \mathcal{N}(L^*) = L(X)^\perp$. Thus $L(V)^\perp = L(X)^\perp$, i.e, $\overline{L(V)} = \overline{L(X)}$. Therefore, there exists $v_n \in V$ such that $Lv_n \to L\rho$ as n tends to infinity. Set $x_n := \rho - v_n$. Then $\|Lx_n\| \to 0$ and $(\rho,x_n) = |\rho|^2$, thus $|(\rho,x_n)|/\,\|Lx_n\| \to \infty$. This ends the proof. □

We will apply this result in the following two subsections to the linear and circular line sources. Both alternatives of this theorem will appear. From this point of view, the linear and the circular line sources behave quite differently.

[4] see Appendix, Definition A.10

1.5.1 The Linear Line Source

In the case where the curve C is the straight line segment of length 2ℓ along the $\hat{e}_3$-axis, the line factor f from (1.50) reduces to

$$f(\theta) \;=\; \int_{-\ell}^{\ell} \psi(s)\, \mathrm{e}^{iks(\cos\theta_0 - \cos\theta)}\, ds\,, \quad \theta \in [0,\pi]\,. \tag{1.56}$$

Note that in this case the factor is independent on ϕ. Therefore, the operator $\mathcal{K}$ can be considered as an operator from $L^2(-\ell,+\ell)$ into $C[0,\pi]$, given by

$$(\mathcal{K}\psi)(\theta) \;:=\; \alpha(\theta) \int_{-\ell}^{\ell} \psi(s)\, \mathrm{e}^{iks(\cos\theta_0 - \cos\theta)}\, ds\,, \quad \theta \in [0,\pi]\,, \tag{1.57}$$

where $\alpha(\theta) \equiv 1$ or $\alpha(\theta) = \sin\theta$. We note that

$$\begin{aligned}\|\mathcal{K}\psi\|^2_{L^2(S^2)} &= \int_0^{2\pi}\int_0^{\pi} \left|(\mathcal{K}\psi)(\theta)\right|^2 \sin\theta\, d\theta\, d\phi \;=\; 2\pi \int_0^{\pi} \left|(\mathcal{K}\psi)(\theta)\right|^2 \sin\theta\, d\theta \\ &= 2\pi \int_{-1}^{1} \left|(K\psi)(t)\right|^2 dt \;=\; 2\pi\, \|K\psi\|^2_{L^2(-1,+1)}\end{aligned}$$

where the operator $K : L^2(-\ell,+\ell) \longrightarrow C[-1,+1]$ is defined by

$$(K\psi)(t) \;:=\; \tilde{\alpha}(t) \int_{-\ell}^{\ell} \psi(s)\, \mathrm{e}^{iks(t_0 - t)}\, ds\,, \quad |t| \leq 1\,, \tag{1.58}$$

where $\tilde{\alpha} \equiv 1$ or $\tilde{\alpha}(t) = \sqrt{1-t^2}$ and $t_0 = \cos\theta_0$. Note that we distinguish between the operators $\mathcal{K} : L^2(-\ell,+\ell) \longrightarrow C(S^2)$ and $K : L^2(-\ell,+\ell) \longrightarrow C[-1,+1]$ in the notation. Analogously, for the noise functional we have

$$\begin{aligned}\|\omega\mathcal{K}\psi\|^2_{L^2(S^2)} &= \int_0^{2\pi}\int_0^{\pi} \omega(\theta,\phi)\, \left|(\mathcal{K}\psi)(\theta)\right|^2 \sin\theta\, d\theta\, d\phi \\ &= 2\pi\, \|\tilde{\omega} K\psi\|^2_{L^2(-1,+1)}\end{aligned}$$

with $\tilde{\omega}(t) := \int_0^{2\pi} \omega(\arccos t, \phi)\, d\phi$, $t \in [-1,1]$.

Therefore, we can express D_ψ, SNR_ψ, and Q_ψ as

$$D_\psi(\theta) = \frac{|(K\psi)(\cos\theta)|^2}{\frac{1}{2}\,\|K\psi\|^2_{L^2(-1,+1)}}, \quad \theta \in [0,\pi], \tag{1.59a}$$

$$SNR_\psi(\theta) = \frac{|(K\psi)(\cos\theta)|^2}{\frac{1}{2}\big\|\tilde{\omega}K\psi\big\|^2_{L^2(-1,+1)}}, \quad \theta \in [0,\pi], \tag{1.59b}$$

$$Q_\psi := \frac{\|\psi\|^2_{L^2(-\ell,\ell)}}{2\pi\,\|K\psi\|^2_{L^2(-1,+1)}}. \tag{1.59c}$$

In the simplest case of a constant current $\psi(s) = 1/(2\ell)$, $|s| \leq \ell$, the factor (1.56) reduces to

$$f(\theta) = \frac{1}{2\ell}\int_{-\ell}^{\ell} \mathrm{e}^{iks(\cos\theta_0 - \cos\theta)}\,ds = \frac{\sin\big[k\ell(\cos\theta_0 - \cos\theta)\big]}{k\ell(\cos\theta_0 - \cos\theta)}, \quad 0 \leq \theta \leq \pi, \tag{1.60}$$

which corresponds to the form (1.7) for the uniform linear array. We observe that both arrays formally coincide if we replace $(2N+1)kd/2$ by $k\ell$ and the denominator of (1.7) by $\left(\frac{2N+1}{2}\right)kd(\cos\theta_0 - \cos\theta)$. This is only a reasonable approximation for small $(2N+1)kd/2$.

If we define γ by $\gamma = \gamma(\theta,\theta_0) = k\ell(\cos\theta_0 - \cos\theta)$ then the plot of $\gamma \mapsto \sin\gamma/\gamma$ in Figure 1.14 is comparable with that in Figure 1.1.

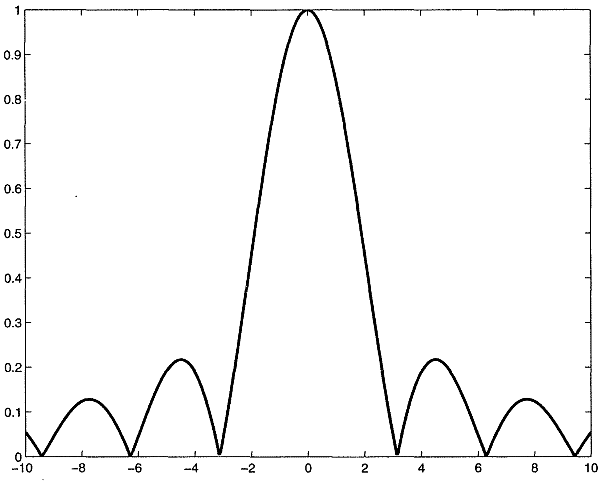


Fig. 1.14. $\gamma \mapsto \left|\frac{\sin\gamma}{\gamma}\right|$

The form of f given by (1.60) is analogous to that of the finite array with the exception that, while there is a main lobe at $\theta = \theta_0$, i.e. $\gamma = 0$, there are no grating lobes, i.e. side lobes of the same magnitude as the main lobe. The beam-width is again given by the angular separation between the first nulls on each side of θ_0. Since the zeros are the roots of the equation

$$k\ell(\cos\theta_0 - \cos\theta) \;=\; j\pi\,, j \in \mathbb{Z}\,,$$

we may compute, for small λ/ℓ, the beam-width just as for the linear array. The approximations are λ/ℓ for the broadside array (i.e. $\theta_0 = 0$) and $2\sqrt{\lambda/\ell}$ for the end-fire array (i.e. $\theta_0 = 0$ or π).

As we show in the next theorem, the operator $K : L^2(-\ell,+\ell) \longrightarrow C[-1,+1]$ satisfies the assumptions of Theorem 1.10.

Theorem 1.11. *The operator $K : L^2(-\ell,+\ell) \longrightarrow C[-1,+1]$, defined in (1.58), is one-to-one and*

$$\sup_{\psi\neq 0} D_\psi(\theta) \;=\; 2\sup_{\psi\neq 0}\frac{|(K\psi)(\cos\theta)|^2}{\|K\psi\|^2_{L^2(-1,+1)}} \;=\; \infty$$

for every $\theta \in [0,\pi]$. If $\alpha(\theta) = \sin\theta$ we have, of course, to assume that $\theta \in (0,\pi)$ since otherwise $D_\psi(\theta) = 0$.

Proof: Without loss of generality we can restrict ourselves to the broadside case $\theta_0 = \pi/2$ since a phase change does not change the injectivity of the operator or the directivity of the array. We will apply Theorem 1.10 to $X = L^2(-\ell,+\ell)$, $Y = L^2(-1,+1)$, and $\rho(s) = \alpha(\theta)\exp(iks\cos\theta)$, $0 \leq s \leq 2\pi$. First we show injectivity. $K\psi = 0$ implies, first, that

$$\int_{-\ell}^{\ell} \psi(s)\,\mathrm{e}^{-ikts}\,ds \;=\; 0 \quad \text{for all } t \in (-1,+1)\,,$$

and, second, by analyticity

$$\int_{-\ell}^{\ell} \psi(s)\,\mathrm{e}^{-its}\,ds \;=\; 0 \quad \text{for all } t \in \mathbb{R}\,.$$

This means, that the Fourier transform of ψ (where ψ has been extended to zero outside $(-\ell,+\ell)$) vanishes. Here, the Fourier transform $\hat{\psi}$ is defined by

$$\hat{\psi}(t) \;=\; \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \psi(s)\,\mathrm{e}^{-ist}\,ds \;=\; \frac{1}{\sqrt{2\pi}}\int_{-\ell}^{\ell} \psi(s)\,\mathrm{e}^{-ist}\,ds\,, \quad t \in \mathbb{R}\,.$$

The injectivity of the Fourier transform implies that also ψ vanishes.

Now we have to show that $\rho \notin \mathcal{R}(K^*)$. Assume, on the contrary, that there exists $g \in L^2(-1,+1)$ with $K^*g = \rho$, i.e.

$$\int_{-1}^{1} \tilde{\alpha}(s)\, g(s)\, \mathrm{e}^{ikst}\, ds \;=\; \rho(t) \;=\; \alpha(\theta)\, \mathrm{e}^{ikt\cos\theta} \quad \text{for all } t \in (-\ell, \ell)\,.$$

By analyticity, this holds for all $t \in \mathbb{R}$. The left hand side is $\sqrt{2\pi}\, \hat{h}(-kt)$ where $\hat{h}$ is the Fourier transform of $h(s) = \begin{cases} \tilde{\alpha}(s)\, g(s)\,, & |s| \le 1\,, \\ 0\,, & |s| > 1\,. \end{cases}$

The Fourier transform of every integrable function decays to zero as $|t|$ tends to infinity. This contradicts the fact that the right hand side $|\rho(t)| = \alpha(\theta) > 0$ for all t. Application of Theorem 1.10 yields the assertion. □

Let us now fix the current ψ and study the the behavior of the directivity when the wave length $\lambda = 2\pi/k$ tends to zero, i.e. k tends to infinity. Again, we can restrict ourselves to the broadside case, i.e. $\theta_0 = \pi/2$. We indicate the dependence of the far field operator K on λ by writing K_λ instead of simply K in (1.58).

From (1.58) we observe that

$$\big(K_\lambda \psi\big)(t) \;=\; \sqrt{2\pi}\, \tilde{\alpha}(t)\, \hat{\psi}_1(kt)$$

where ψ_1 denotes the extension of ψ by zero onto $\mathbb{R}$ and $\hat{\psi}_1$ its Fourier transform. As before, we compute the norm of $K_\lambda \psi$:

$$\begin{aligned} \|K_\lambda \psi\|^2_{L^2(-1,+1)} &= 2\pi \int_{-1}^{1} \tilde{\alpha}(t)^2 \left|\hat{\psi}_1(kt)\right|^2 dt \\ &= \frac{2\pi}{k} \int_{-k}^{k} \tilde{\alpha}(t/k)^2 \left|\hat{\psi}_1(t)\right|^2 dt \;=\; \lambda \int_{-\infty}^{\infty} \rho_\lambda(t)^2 \left|\hat{\psi}_1(t)\right|^2 dt \\ &= \lambda \left\|\rho_\lambda \hat{\psi}_1\right\|^2_{L^2(\mathbb{R})} \end{aligned}$$

where $\rho_\lambda(t) = \tilde{\alpha}(t/k)$ for $|t| \le k = 2\pi/\lambda$ and zero otherwise. We note that $\rho_\lambda(t)$ tends to one as $\lambda \to 0$ for every $t \in \mathbb{R}$. Furthermore, $0 \le \rho_\lambda \le 1$. From Lebesgue's Theorem on dominated convergence, we conclude that $\left\|\rho_\lambda \hat{\psi}_1\right\|^2_{L^2(\mathbb{R})} \to \left\|\hat{\psi}_1\right\|^2_{L^2(\mathbb{R})}$ as λ tends to zero and thus, by Plancherel's Theorem,

$$\frac{1}{\lambda}\, \|K_\lambda \psi\|^2_{L^2(-1,+1)} \;\longrightarrow\; \left\|\hat{\psi}_1\right\|^2_{L^2(\mathbb{R})} \;=\; \|\psi\|^2_{L^2(-\ell,\ell)}\,, \quad \lambda \to 0\,.$$

Therefore, $\frac{\lambda}{4\ell}\, D_{\lambda,\psi}(\pi/2)$ in the broadside direction is

$$\frac{\lambda}{4\ell} D_{\lambda,\psi}(\pi/2) = \frac{\frac{1}{2\ell}\,\left|(K_\lambda\psi)(0)\right|^2}{\frac{1}{\lambda}\,\|K_\lambda\psi\|^2_{L^2(-1,+1)}} = \frac{\frac{1}{2\ell}\,\left|\int_{-\ell}^{\ell}\psi(s)\,ds\right|^2}{\frac{1}{\lambda}\,\|K_\lambda\psi\|^2_{L^2(-1,+1)}} \tag{1.61}$$

and the right hand side converges to

$$D_\infty(\psi) := \lim_{\lambda\to 0}\left[\frac{\lambda}{4\ell} D_{\lambda,\psi}(\pi/2)\right] = \frac{\frac{1}{2\ell}\,\left|\int_{-\ell}^{\ell}\psi(s)\,ds\right|^2}{\|\psi\|^2_{L^2(-\ell,\ell)}}. \tag{1.62}$$

Notice that, by the Cauchy-Schwarz inequality, $D_\infty(\psi) \leq 1$ for every $\psi \in L^2(-\ell,\ell)$ and equality holds if and only if the current ψ is constant. As we have seen in Theorem 1.11, for a finite line source and positive wave length $\lambda > 0$, it is possible to increase the specific gain $\frac{\lambda}{4\ell} D_{\lambda,\psi}(\pi/2)$ over that for a constant current. However, the quantity $D_\infty(\psi)$ must at the same time decrease. To quote Taylor ([133]):

> "It is in this sense, and only in this sense, that a uniform distribution has a specific gain $\left[\frac{\lambda}{4\ell} D_{\lambda,\psi}(\pi/2)\right]$ greater than that of any other type of line-source distribution."

Reference to the expression (1.62) shows that the directivity $D_{\lambda,\psi}(\pi/2)$ tends to infinity in the broadside direction:

$$D_{\lambda,\psi}(\pi/2) = \frac{4\ell}{\lambda}\left[D_\infty(\psi) + o(1)\right], \quad \lambda \to 0.$$

For directions $\theta \neq \pi/2$ and sufficiently smooth ψ, however, the directivities $D_{\lambda,\psi}(\theta)$ tend to zero as λ tends to zero. Indeed, the numerator $(K_\lambda\psi)(\cos\theta) = \alpha(\theta)\int_{-\ell}^{\ell}\psi(s)\exp(-iks\cos\theta)\,ds$ tends to zero of order $1/k$ as k tends to infinity. This is easily seen for continuously differentiable functions ψ using integration by parts:

$$\int\limits_{-\ell}^{\ell}\psi(s)\,\exp(-iks\cos\theta)\,ds = \frac{i}{k\cos\theta}\left[\psi(s)\,\exp(-iks\cos\theta)|_{-\ell}^{+\ell}\right.$$
$$\left. - \int\limits_{-\ell}^{\ell}\psi'(s)\,\exp(-iks\cos\theta)\,ds\right] = \mathcal{O}(1/k),$$

and thus $\left|(K_\lambda\psi)(\cos\theta)\right|^2 = \mathcal{O}(\lambda^2)$.

Following Taylor, we may introduce the super-gain ratio as a measure of the relative growth of these quantities.

Definition 1.12. *For a linear line source with current $\psi \in L^2(-\ell,\ell)$, $\psi \neq 0$, we define the* **super-gain ratio** *$\gamma_\lambda(\psi)$ by*

$$\gamma_\lambda(\psi) := \frac{\frac{\lambda}{4\ell} D_{\lambda,\psi}(\pi/2)}{D_\infty(\psi)} = \lambda\frac{\|\psi\|^2_{L^2(-\ell,\ell)}}{\|K_\lambda\psi\|^2_{L^2(-1,+1)}}. \tag{1.63}$$

We note that the super-gain ratio $\gamma_\lambda(\psi)$ coincides with the Q–factor (1.59c) up to the factor $2\pi\lambda$.

We illustrate this with a simple example.

Example 1.13. For a uniform distribution $\psi = 1$ we compute $\|\psi\|^2_{L^2(-\ell,\ell)} = 2\ell$ and

$$(K_\lambda\psi)(t) \;=\; \int\limits_{-\ell}^{\ell} \mathrm{e}^{-ikst}\,ds \;=\; 2\,\frac{\sin(k\ell t)}{kt}\,.$$

Thus, using $\sin^2 x = \frac{1}{2}\big(1 - \cos(2x)\big)$ and partial integration,

$$\begin{aligned}
\|K_\lambda\psi\|^2_{L^2(-1,+1)} &= \frac{4}{k^2}\int\limits_{-1}^{1}\frac{\sin^2(k\ell t)}{t^2}\,dt \;=\; \frac{8}{k^2}\int\limits_{0}^{1}\frac{\sin^2(k\ell t)}{t^2}\,dt \\
&= \frac{4}{k^2}\int\limits_{0}^{1}\frac{1-\cos(2k\ell t)}{t^2}\,dt \\
&= \frac{4}{k^2}\big[\cos(2k\ell) - 1 + 2k\ell\,\mathit{Si}(2k\ell)\big] \\
&= \frac{8\ell}{k}\left[\mathit{Si}(2k\ell) \;-\; \frac{\sin^2(k\ell)}{k\ell}\right],
\end{aligned}$$

where the sine-integral $\mathit{Si}(x)$ is defined by

$$\mathit{Si}(x) \;:=\; \int\limits_{0}^{x}\frac{\sin s}{s}\,ds\,, \quad x > 0\,. \tag{1.64}$$

The super-gain ratio $\gamma_\lambda(1)$ for the constant current $\psi = 1$ is therefore given as

$$\gamma_\lambda(1) \;=\; \lambda\,\frac{2\ell}{\frac{8\ell}{k}\left[\mathit{Si}(2k\ell) \;-\; \frac{\sin^2(k\ell)}{k\ell}\right]} \;=\; \frac{\pi/2}{\mathit{Si}(2k\ell) \;-\; \frac{\sin^2(k\ell)}{k\ell}}\,. \tag{1.65}$$

This example motivates the following definition:

Definition 1.14. *Let the super-gain ratio $\gamma_\lambda(\psi)$ for a linear line source and wave length λ be defined by (1.63). A* **super-gain line source** *is one for which $\gamma_\lambda(\psi) > \gamma_\lambda(1)$ where $\gamma_\lambda(1)$ is given by (1.65).*

The effects of super-gain (see [115]) are that the pattern becomes very large in the invisible region while changing little in the visible. Likewise, the aperture distribution oscillates rapidly over the aperture. Both of these effects severely limit the effectiveness of the antenna.

1.5.2 The Circular Line Source

Analogous results can be derived for the circular line source much as before. We assume the circular line source occupying the circle C in the (x_1, x_2)−plane and given by $\boldsymbol{y}(s) = a\big(\cos s, \sin s, 0\big)^\top \in \mathbb{R}^3$, $0 \leq s \leq 2\pi$. The operator $\mathcal{K} : L^2(0, 2\pi) \longrightarrow C(S^2)$ for this particular antenna and $\alpha(\hat{\boldsymbol{x}}) \equiv 1$ is given by

$$\big(\mathcal{K}\psi\big)(\hat{\boldsymbol{x}}) := a \int_0^{2\pi} \psi(s)\, \mathrm{e}^{ik\boldsymbol{y}(s)\cdot(\hat{\boldsymbol{x}}_0 - \hat{\boldsymbol{x}})}\, ds\,, \quad \hat{\boldsymbol{x}} \in S^2\,. \tag{1.66}$$

Analogously to Theorem 1.11 we prove

Theorem 1.15. *The operator $\mathcal{K} : L^2(0, 2\pi) \longrightarrow C(S^2)$, defined in (1.66), is one-to-one and*

$$\sup_{\psi \neq 0} D_\psi(\hat{\boldsymbol{x}}) = \sup_{\psi \neq 0} \frac{|(\mathcal{K}\psi)(\hat{\boldsymbol{x}})|^2}{\frac{1}{4\pi}\, \|\mathcal{K}\psi\|^2_{L^2(S^2)}} < \infty$$

if and only if $\theta \neq \pi/2$ where (ϕ, θ) denote the polar coordinates of $\hat{\boldsymbol{x}}$.

We note the difference of this result to the case of a linear line source. While for the linear line source the directivity can be arbitrarily large, for the circular line source it remains bounded for all directions except those in the plane of the array.

Proof: Again, we can assume that $\hat{\boldsymbol{x}}_0 = \hat{\mathbf{e}}_3$. Again, we begin with a proof of injectivity. Using spherical polar coordinates and $\boldsymbol{y}(s) \cdot \hat{\boldsymbol{x}} = a \sin\theta\, \cos(s - \phi)$, we write $\mathcal{K}\psi$ as

$$\big(\mathcal{K}\psi\big)(\hat{\boldsymbol{x}}) = \big(\mathcal{K}\psi\big)(\theta, \phi) = a \int_0^{2\pi} \psi(s)\, \mathrm{e}^{-ika\sin\theta\cos(s-\phi)}\, ds\,, \tag{1.67}$$

for $\theta \in [0, \pi]$, $\phi \in [0, 2\pi]$. The Jacobi-Anger expansion yields

$$\big(\mathcal{K}\psi\big)(\theta, \phi) = a \sum_{n \in \mathbb{Z}} (-i)^n\, J_n(ka \sin\theta)\, \mathrm{e}^{in\phi} \int_0^{2\pi} \psi(s)\, \mathrm{e}^{-ins}\, ds\,.$$

For fixed θ, this is a Fourier series with respect to $\phi \in [0, 2\pi]$. Let $\mathcal{K}\psi = 0$ on S^2. We choose $\theta > 0$ so small such that $J_n(ka \sin\theta) \neq 0$ for all $n \in \mathbb{Z}$. From $(\mathcal{K}\psi)(\theta, \phi) = 0$ for all $\phi \in [0, 2\pi]$ we conclude that all the Fourier coefficients $\int_0^{2\pi} \psi(s)\, \mathrm{e}^{-ins}\, ds$ of ψ vanish, i.e. $\psi = 0$.

Again, we will apply Theorem 1.10 to $\mathcal{K} : L^2(0, 2\pi) \longrightarrow L^2(S^2)$, given by (1.66). We note that $(\mathcal{K}\psi)(\hat{\boldsymbol{x}}) = (\psi, \rho)_{L^2(0,2\pi)}$ where

$$\rho(t) = a\, \mathrm{e}^{ik\, \boldsymbol{y}(t)\cdot\hat{\boldsymbol{x}}} = a\, \mathrm{e}^{ika\sin\theta\cos(t-\phi)} = a \sum_{n \in \mathbb{Z}} i^n\, J_n(ka \sin\theta)\, \mathrm{e}^{in(t-\phi)}\,,$$

$0 \le t \le 2\pi$. We have to study the solvability of the equation $\mathcal{K}^* g = \rho$ in $L^2(S^2)$. The operator $\mathcal{K}^* : L^2(S^2) \longrightarrow L^2(0, 2\pi)$ is given by

$$\big(\mathcal{K}^* g\big)(t) \;=\; a \int\limits_{S^2} g(\hat{\boldsymbol{x}})\, \mathrm{e}^{ik\boldsymbol{y}(t)\cdot\hat{\boldsymbol{x}}}\, dS(\hat{\boldsymbol{x}})\,, \quad t \in [0, 2\pi]\,.$$

Let us first consider the case $\theta \ne \pi/2$, i.e. $\sin\theta < 1$. We make an ansatz for g in the form

$$g(\hat{\boldsymbol{x}}) \;=\; g(\theta', \phi') \;=\; |\cos\theta'| \sum_{n\in\mathbb{Z}} g_n \,(\sin\theta')^{|n|}\, \mathrm{e}^{in\phi'} \quad \text{with some } g_n \in \mathbb{C}\,. \tag{1.68}$$

Then, using $\boldsymbol{y}(t)\cdot\hat{\boldsymbol{x}} = a\sin\theta'\,\cos(t-\phi')$ and the Jacobi Anger expansion again,

$$\begin{aligned}
&(\mathcal{K}^* g)(t)\\
&= \sum_{n\in\mathbb{Z}} g_n \int\limits_0^{2\pi}\int\limits_0^{\pi} \mathrm{e}^{ika\sin\theta'\cos(t-\phi')}\,\mathrm{e}^{in\phi'}\,(\sin\theta')^{|n|+1}\,|\cos\theta'|\,d\theta'\,d\phi'\\
&= \sum_{n,m\in\mathbb{Z}} g_n\, i^m \int\limits_0^{2\pi} \mathrm{e}^{im(t-\phi')}\,\mathrm{e}^{in\phi'}\,d\phi' \int\limits_0^{\pi} J_m(ka\sin\theta')\,(\sin\theta')^{|n|+1}\,|\cos\theta'|\,d\theta'\\
&= 2\pi \sum_{n\in\mathbb{Z}} g_n\, i^n\, \mathrm{e}^{int} \int\limits_0^{\pi} J_n(ka\sin\theta')\,(\sin\theta')^{|n|+1}\,|\cos\theta'|\,d\theta'
\end{aligned}$$

The integral is well known as Sonine's first finite integral (see [139], p. 373) and its value is

$$\begin{aligned}
\int\limits_0^{\pi} J_n(ka\sin\theta')\,(\sin\theta')^{|n|+1}\,|\cos\theta'|\,d\theta' &= 2\int\limits_0^{\pi/2} J_n(ka\sin\theta')\,(\sin\theta')^{n+1}\,\cos\theta'\,d\theta'\\
&= \frac{2}{ka}\,J_{n+1}(ka)\,, \quad n \ge 0\,.
\end{aligned}$$

For $n < 0$ we use $J_n(t) = (-1)^n J_{-n}(t)$ and get an analogous formula.

Let us assume for the moment that ka is such that $J_n(ka) \ne 0$ for all $n \in \mathbb{Z}$. Then we can solve for g_n and have

$$g_n \;=\; \frac{a\,J_{|n|}(ka\sin\theta)}{2\pi\,\frac{2}{ka}\,J_{|n|+1}(ka)}\,\mathrm{e}^{-in\phi}\,, \quad n \in \mathbb{Z}\,.$$

It remains to show that $g \in L^2(S^2)$ i.e., $\sum_{n\in\mathbb{Z}} |g_n|^2 < \infty$. This follows from the asymptotic behaviour of the Bessel functions

$$J_n(t) \;=\; \frac{1}{n!}\left(\frac{t}{2}\right)^n \big[1 \;+\; \mathcal{O}(1/n)\big]\,, \quad n \to \infty\,,$$

uniformly with respect to t in compact subsets of $\mathbb{R}$. Indeed,

$$|g_n| \;=\; \frac{a}{2\pi}\,|n+1|\,(\sin\theta)^{|n|}\left[1 \;+\; \mathcal{O}(1/|n|)\right], \quad n \to \pm\infty,$$

and $\sin\theta < 1$.

If $J_{n_0}(ka) = 0$ for some $n_0 \in \mathbb{Z}$ then we change the factor of g_{n_0} in the assumed form (1.68) of g to $|\cos\theta'|\,(\sin\theta')^{|n_0|}\,\epsilon(\theta')\,\mathrm{e}^{in_0\phi'}$ where ϵ is chosen such that

$$\int_0^{\pi} J_{n_0}(ka\sin\theta')\,(\sin\theta')^{|n_0|+1}\,\epsilon(\theta')\,|\cos\theta'|\,d\theta' \;\neq\; 0\,.$$

This does not effect the convergence property of the sequence $\{g_n\}_{n=1}^{\infty}$.

Let us now consider the case $\theta = \pi/2$ i.e., $\sin\theta = 1$. Assume, that $K^*g = \rho$ were solvable in $L^2(S^2)$. Then, as before,

$$\int_0^{2\pi}\int_0^{\pi} \mathrm{e}^{ika\sin\theta'\cos(t-\phi')}\,g(\theta',\phi')\,\sin\theta'\,d\theta'\,d\phi' \;=\; a\,\mathrm{e}^{ika\cos(t-\phi)}\,, \quad 0 \le t \le 2\pi\,,$$

i.e. with the Jacobi-Anger expansion,

$$\sum_{n\in\mathbb{Z}} i^n \int_0^{\pi} J_n(ka\sin\theta') \int_0^{2\pi} \mathrm{e}^{-in\phi'}\,g(\theta',\phi')\,d\phi'\,\sin\theta' d\theta'\,\mathrm{e}^{int} \;=$$

$$a\sum_{n\in\mathbb{Z}} i^n\,J_n(ka)\,\mathrm{e}^{in(t-\phi)} \quad \text{for all } t \in [0,2\pi]\,.$$

Equating coefficients, we get

$$\int_0^{\pi} J_n(ka\sin\theta')\,\sin\theta' \int_0^{2\pi} \mathrm{e}^{-in\phi'}\,g(\theta',\phi')\,d\phi'\,d\theta' \;=\; a\,J_n(ka)\,\mathrm{e}^{-in\phi}$$

for all $n \in \mathbb{Z}$. Multiplying this equation by $n!\,\left(\frac{2}{ka}\right)^n \exp(in\phi)$ yields

$$\int_0^{\pi} f_n(\theta')\,d\theta' \;=\; n!\left(\frac{2}{ka}\right)^n a\,J_n(ka) \;=\; a\left[1 \;+\; \mathcal{O}(1/n)\right], \quad n \to \infty, \quad (1.69)$$

where

$$f_n(\theta') \;=\; n!\left(\frac{2}{ka}\right)^n \mathrm{e}^{in\phi}\,J_n(ka\sin\theta')\,\sin\theta' \int_0^{2\pi} \mathrm{e}^{-in\phi'}\,g(\theta',\phi')\,d\phi'\,.$$

The functions f_n are bounded uniformly in n and $f_n(\theta') \to 0$, $n \to \infty$, for all $\theta' \neq \pi/2$ by the asymptotic form of J_n. Lebesgue's theorem on dominated convergence yields $\int_0^\pi f_n(\theta')\,d\theta' \to 0$, a contradiction to (1.69). This proves the theorem. □

Just as in our previous discussion, we study the asymptotic form as λ tends to zero, i.e. as k tends to infinity. We write again $\mathcal{K}_\lambda$ instead of $\mathcal{K}$ to indicate the dependence on λ. We first prove a special case of the **method of stationary phase** that will allow us to compute the asymptotic form of the norm $\|K_\lambda\psi\|_{L^2(S^2)}$:

Lemma 1.16. *Let $\varphi \in C^2[a,b]$ be strictly monotonic with $\varphi'(s) > 0$ for all $s \in [a,b)$ (or $\varphi'(s) < 0$ for all $s \in [a,b)$) and $\varphi(a) = 0$. Let $g \in C^1[a,b]$ and*

$$I_k \;:=\; \int\limits_a^b g(s)\,\frac{\sin k\varphi(s)}{\varphi(s)}\,ds\,, \quad k > 0\,.$$

Then

$$I_k \;\longrightarrow\; \frac{\pi g(a)}{2\,|\varphi'(a)|} \quad \text{as } k \to \infty\,.$$

The same assertion holds if the roles of a and b are interchanged.

We add a proof for the convenience of the reader.

Proof: Let $\varphi' > 0$ on $[a,b)$. By the change of variables $s' = s - a$ we can assume that $a = 0$. We make the substitution $\sigma = \varphi(s)$, and set $\sigma_1 := \varphi(b)$ which yields

$$I_k \;=\; \int\limits_0^{\sigma_1} h(\sigma)\,\frac{\sin(k\sigma)}{\sigma}\,d\sigma \quad \text{where} \quad h(\sigma) \;:=\; \frac{g\big(\varphi^{-1}(\sigma)\big)}{\varphi'\big(\varphi^{-1}(\sigma)\big)}\,, \quad 0 \le \sigma < \sigma_1\,.$$

We note that our assumptions on g and φ guarantee that $h \in C^1[0,\sigma_1)$ and $h \in L^1(0,\sigma_1)$. We begin by assuming that $h \in C^1[0,\sigma_1]$ rather than merely in $C^1[0,\sigma_1)$. Then we write

$$I_k \;=\; h(0)\int\limits_0^{\sigma_1} \frac{\sin(k\sigma)}{\sigma}\,d\sigma \;+\; \int\limits_0^{\sigma_1} [h(\sigma) - h(0)]\,\frac{\sin(k\sigma)}{\sigma}\,d\sigma\,, \tag{1.70}$$

and consider the integrals separately. The change of variables $s = k\sigma$ in the first integral yields

$$\int\limits_0^{\sigma_1} \frac{\sin(k\sigma)}{\sigma}\,d\sigma \;=\; \int\limits_0^{k\sigma_1} \frac{\sin s}{s}\,ds \;=\; \mathit{Si}(k\sigma_1)$$

where the sine-integral $\mathit{Si}(x)$ is defined in (1.64). From $\lim_{x\to\infty} \mathit{Si}(x) = \pi/2$ (see [90]) we conclude that the first term of (1.70) tends to $h(0)\pi/2 = g(0)\pi/(2\varphi'(0))$ as k tends to infinity.

It remains to show that the second term in (1.70) vanishes as k tends to infinity. To this end we use the fundamental theorem of calculus and integration by parts:

$$\begin{aligned}\int_0^{\sigma_1} [h(\sigma) - h(0)] \frac{\sin(k\sigma)}{\sigma}\, d\sigma &= \int_0^{\sigma_1}\int_0^{\sigma} h'(\tau) \frac{\sin(k\sigma)}{\sigma}\, d\tau\, d\sigma \\ &= \int_0^{\sigma_1} h'(\tau) \int_\tau^{\sigma_1} \frac{\sin(k\sigma)}{\sigma}\, d\sigma\, d\tau \\ &= \int_0^{\sigma_1} h'(\tau) \big[\mathit{Si}(k\sigma_1) - \mathit{Si}(k\tau)\big]\, d\tau\,.\end{aligned}$$

To this expression we apply Lebesgue's theorem on dominated convergence. Since $\mathit{Si}(k\sigma_1) - \mathit{Si}(k\tau)$ tends to zero as $k \to \infty$ for every $\tau > 0$ and is uniformly bounded by $2\,\|\mathit{Si}\|_{C[0,\infty)}$ we conclude that this integral tends to zero as k tends to infinity, and the proof is complete for smooth h.

Let us now drop the simplifying assumption that $h \in C^1[0,\sigma_1]$ and consider the general case where merely $h \in C^1[0,\sigma_1) \cap L^1(0,\sigma_1)$. Let $\epsilon > 0$ be arbitrary. Choose $\tilde{h} \in C^1[0,\sigma_1]$ with $\tilde{h} = h$ on $[0,\sigma_1/2]$ and $\int_0^{\sigma_1} |\tilde{h} - h| ds \le \sigma_1\epsilon/4$. Application of the above arguments to $\tilde{h}$ yields

$$\tilde{I}_k := \int_0^{\sigma_1} \tilde{h}(\sigma) \frac{\sin(k\sigma)}{\sigma}\, d\sigma \longrightarrow \tilde{h}(0)\, \frac{\pi}{2} = \frac{\pi g(0)}{2\varphi'(0)}\,.$$

Furthermore, we can make the estimates

$$\begin{aligned}\left| I_k - \frac{\pi g(0)}{2\varphi'(0)} \right| &\le \left| I_k - \tilde{I}_k \right| + \left| \tilde{I}_k - \frac{\pi g(0)}{2\varphi'(0)} \right| \\ &\le \int_{\sigma_1/2}^{\sigma_1} \frac{|\tilde{h}(\sigma) - h(\sigma)|}{\sigma}\, d\sigma + \left| \tilde{I}_k - \frac{\pi g(0)}{2\varphi'(0)} \right| \\ &\le \frac{2}{\sigma_1} \int_{\sigma_1/2}^{\sigma_1} |\tilde{h}(\sigma) - h(\sigma)|\, d\sigma + \left| \tilde{I}_k - \frac{\pi g(0)}{2\varphi'(0)} \right| \\ &\le \frac{\epsilon}{2} + \left| \tilde{I}_k - \frac{\pi g(0)}{2\varphi'(0)} \right|,\end{aligned}$$

this last quantity being less than ϵ for sufficiently large k. This proves the assertion of the lemma for the case that $\varphi' > 0$ on $[a,b)$. In the case of $\varphi' < 0$ on $[a,b)$ we replace φ by $-\varphi$. If $\varphi(b) = 0$ instead of $\varphi(a) = 0$ we use the substitution $s = b + s'(a-b)$, $s' \in [0,1]$, which maps the zero into the left endpoint. This completes the proof. □

Having this lemma at our disposal, we may now establish the asymptotic behaviour of the circular line factor $\mathcal{K}_\lambda\psi$ from (1.66) for continuously differentiable functions ψ and $\alpha(\hat{\boldsymbol{x}}) \equiv 1$. Since

$$|\mathcal{K}_\lambda\psi(\hat{\boldsymbol{x}})|^2 = a^2 \int_0^{2\pi}\int_0^{2\pi} \psi(s)\,\overline{\psi(t)}\,\mathrm{e}^{ik(\boldsymbol{y}(t)-\boldsymbol{y}(s))\cdot\hat{\boldsymbol{x}}}\,ds\,dt\,,$$

the norm of the element $K_\lambda(\psi)$ for this case can be computed explicitly.

$$\begin{aligned}
\|\mathcal{K}_\lambda\psi\|^2_{L^2(S^2)} &= \int_{S^2} |K_\lambda\psi(\hat{\boldsymbol{x}})|^2\,dS(\hat{\boldsymbol{x}}) \\
&= a^2 \int_0^{2\pi}\int_0^{2\pi} \psi(s)\,\overline{\psi(t)} \int_{S^2} \mathrm{e}^{ik(\boldsymbol{y}(t)-\boldsymbol{y}(s))\cdot\hat{\boldsymbol{x}}}\,ds(\hat{\boldsymbol{x}})\,dS\,dt \\
&= a^2 \int_0^{2\pi}\int_0^{2\pi} \psi(s)\,\overline{\psi(t)} \int_{S^2} \mathrm{e}^{ik|\boldsymbol{y}(t)-\boldsymbol{y}(s)|\cos\theta}\,dS(\hat{\boldsymbol{x}})\,ds\,dt \\
&= 2\pi a^2 \int_0^{2\pi}\int_0^{2\pi} \psi(s)\,\overline{\psi(t)} \int_0^{\pi} \mathrm{e}^{ik|\boldsymbol{y}(t)-\boldsymbol{y}(s)|\cos\theta}\,\sin\theta\,d\theta\,ds\,dt \\
&= 2\pi a^2 \int_0^{2\pi}\int_0^{2\pi} \psi(s)\,\overline{\psi(t)} \int_{-1}^{1} \mathrm{e}^{ik|\boldsymbol{y}(t)-\boldsymbol{y}(s)|\tau}\,d\tau\,ds\,dt \\
&= 4\pi a^2 \int_0^{2\pi}\int_0^{2\pi} \psi(s)\,\overline{\psi(t)}\,\frac{\sin[k\,|\boldsymbol{y}(t)-\boldsymbol{y}(s)|]}{k\,|\boldsymbol{y}(t)-\boldsymbol{y}(s)|}\,ds\,dt\,.
\end{aligned}$$

With $|\boldsymbol{y}(t)-\boldsymbol{y}(s)|^2 = 2a^2 - 2a^2\cos(s-t) = 4a^2\sin^2\frac{s-t}{2}$ this yields

$$\|\mathcal{K}_\lambda\psi\|^2_{L^2(S^2)} = \frac{4\pi a^2}{k}\int_Q \psi(s)\,\overline{\psi(t)}\,\frac{\sin\left[k\varphi\left(\frac{s-t}{2}\right)\right]}{\varphi\left(\frac{s-t}{2}\right)}\,d(s,t)$$

where $\varphi(\sigma) = 2a\sin\sigma$ and $Q = [0,2\pi]^2 \subset \mathbb{R}^2$. Now we transform this integral by setting $\frac{s-t}{2} = \sigma$ and $\frac{s+t}{2} = \tau$. Noting that the determinant of this transform is $1/2$, this yields

$$\begin{aligned}
\|\mathcal{K}_\lambda\psi\|^2_{L^2(S^2)} &= \frac{8\pi a^2}{k}\int_{Q'} \psi(\tau+\sigma)\,\overline{\psi(\tau-\sigma)}\,\frac{\sin[k\varphi(\sigma)]}{\varphi(\sigma)}\,d(\sigma,\tau) \\
&= \frac{8\pi a^2}{k}\int_{-\pi}^{\pi} g(\sigma)\,\frac{\sin[k\varphi(\sigma)]}{\varphi(\sigma)}\,d\sigma
\end{aligned}$$

where

$$g(\sigma) \;=\; \int\limits_{|\sigma|}^{2\pi-|\sigma|} \psi(\tau+\sigma)\,\overline{\psi(\tau-\sigma)}\,d\tau\,, \quad |\sigma| \le \pi\,.$$

In the interval $[-\pi, \pi]$ the function φ has zeros in $0, \pm\pi$. Therefore, we split the interval $[-\pi, \pi]$ of integration into $[-\pi, -\pi/2] \cup [-\pi/2, 0] \cup [0, \pi/2] \cup [\pi/2, \pi]$ and apply the previous lemma in each of these intervals. From $g(0) = \|\psi\|^2_{L^2(0,2\pi)}$ and $g(\pm\pi) = 0$ and $\varphi'(0) = 2a$ we conclude that

$$\int\limits_{-\pi}^{\pi} g(\sigma)\,\frac{\sin[k\varphi(\sigma)]}{\varphi(\sigma)}\,d\sigma \;\longrightarrow\; 2\,\frac{\|\psi\|^2_{L^2(0,2\pi)}}{2a}\,\frac{\pi}{2} \;=\; \frac{\pi\,\|\psi\|^2_{L^2(0,2\pi)}}{2a} \quad \text{as } k\to\infty\,,$$

and thus that

$$\frac{1}{\lambda}\,\|\mathcal{K}_\lambda\psi\|^2_{L^2(S^2)} \;=\; \frac{k}{2\pi}\,\|\mathcal{K}_\lambda\psi\|^2_{L^2(S^2)} \;\longrightarrow\; 2\pi a\,\|\psi\|^2_{L^2(0,2\pi)}\,, \quad \lambda\to 0\,.$$

As above we consider $\frac{\lambda}{4\pi a}\,D_{\lambda,\psi}(\hat{e}_3)$ and observe that

$$\frac{\lambda}{4\pi a}\,D_{\lambda,\psi}(\hat{e}_3) \;=\; \frac{\frac{1}{2\pi}\left|\int_0^{2\pi}\psi(s)\,ds\right|^2}{\frac{1}{2\pi a\lambda}\,\|\mathcal{K}_\lambda\psi\|^2_{L^2(S^2)}} \tag{1.71}$$

which converges to

$$D_\infty(\psi) \;:=\; \lim_{\lambda\to 0}\left[\frac{\lambda}{4\pi a}\,D_{\lambda,\psi}(\hat{e}_3)\right] \;=\; \frac{\frac{1}{2\pi}\left|\int_0^{2\pi}\psi(s)\,ds\right|^2}{\|\psi\|^2_{L^2(0,2\pi)}}\,. \tag{1.72}$$

Again, we observe that, by the Cauchy-Schwarz inequality, $D_\infty(\psi) \le 1$ for every
$\psi \in L^2(0, 2\pi)$ and equality holds if and only if the current ψ is constant.

For the circular loop, the super-gain ratio $\gamma_\lambda(\psi)$ then takes the form

$$\gamma_\lambda(\psi) \;:=\; \frac{\frac{\lambda}{4\pi a}\,D_{\lambda,\psi}(\hat{e}_3)}{D_\infty(\psi)} \;=\; 2\pi\lambda\,a\,\frac{\|\psi\|^2_{L^2(0,2\pi)}}{\|\mathcal{K}_\lambda\psi\|^2_{L^2(S^2)}}\,. \tag{1.73}$$

We note that again the super-gain ratio coincides with the Q–factor (1.54b) up to the factor $2\pi\lambda$ (since $\|\psi\|^2_{L^2(C)} = a\,\|\psi\|^2_{L^2(0,2\pi)}$ for the circular line source of radius a).

Along with the desire to control super-gain, there are other constraints that arise from power considerations. In the final part of this section, we illustrate their importance in the analysis of optimization problems in antenna theory.

Considerable experience has led to an understanding that certain physical limitations play an important role in the design of efficient antennas.

For example, we have seen that there is a theoretical maximum value for the relative directivity of an array with a given element configuration. However, this theoretical maximal value may be attained only at the expense of very low values of the efficiency index, a quantity which was introduced in Section 1.3 (formula (1.30)). Recall that the efficiency index, as defined there, is a measure of the amount of power radiated into the far field as a portion of the power at the "surface" of the antenna. Evidently, in order to measure the lack of efficiency, some arbitrary standard must be chosen against which comparisons will be made. Such a standard may be, for example, the value of the directivity of a uniformly fed array and its corresponding radiation efficiency.

1.5.3 Numerical Quadrature

Usually, for more complicated geometries or current distributions it is not possible to compute the line factors analytically. Therefore, one is led to the problem how to compute the factors numerically. Note that the line factors are of the form

$$I(f) \;:=\; \int\limits_a^b f(s)\, ds \tag{1.74}$$

for the function $f : [a,b] \longrightarrow \mathbb{C}$, defined by $f(s) = \psi(s)\, e^{ik\boldsymbol{y}(s)\cdot(\hat{\boldsymbol{x}}_0 - \hat{\boldsymbol{x}})}$, $s \in [a,b]$. For the numerical evaluation of (1.74) one replaces the integral $I(f)$ by weighted sums of the form

$$\sum_{j=1}^{n} w_j\, f(t_j)$$

with weights $w_j \in \mathbb{R}$ and node points $t_j \in [a,b]$. Since both depend on n we often indicate this dependence and write $w_j^{(n)}$ and $t_j^{(n)}$, respectively. One is then interested in the problem how well the sum

$$Q_n(f) \;:=\; \sum_{j=1}^{n} w_j^{(n)}\, f\big(t_j^{(n)}\big) \tag{1.75}$$

approximates the integral $I(f)$. In general, this depends on the choices of $w_j^{(n)}$, $t_j^{(n)}$, and the smoothness of the function f. For a rigorous and comprehensive overview we refer to Davis and Rabinowitz [32]. In this subsection we will only state some theorems which are of particular importance to us.

First, we consider the case of a circular line source or, more generally, a closed loop antenna of arbitrary shape. In this case, we can parametrize the closed curve C by a 2π-periodic function $\boldsymbol{y} : [0, 2\pi] \longrightarrow \mathbb{R}^3$. This leads to the computation of integrals of the form

$$I(f) \;=\; \int\limits_0^{2\pi} f(s)\,ds \tag{1.76}$$

with $2\pi-$periodic functions f. For these integrals one can show that (in a certain sense) the composite **trapezoidal rule** is the best possible. In particular one can show:

Theorem 1.17. *Let $N \in \mathbb{N}$ and define $t_j = j\frac{\pi}{N}$ for $j = 0, \ldots, 2N-1$ and*

$$T_N(f) \;:=\; \frac{\pi}{N} \sum_{j=0}^{2N-1} f(t_j) \tag{1.77}$$

for any $2\pi-$periodic and continuous function $f : \mathbb{R} \longrightarrow \mathbb{C}$. Then we have:

(a) If $f \in C^p(\mathbb{R})$ for some odd $p \in \mathbb{N}$ is $2\pi-$periodic then there exists $c = c(p) > 0$ with

$$\bigl|I(f) - T_N(f)\bigr| \;\le\; \frac{c}{N^p} \int\limits_0^{2\pi} \bigl|f^{(p)}(s)\bigr|\,ds\,. \tag{1.78a}$$

(b) If $f : \mathbb{R} \longrightarrow \mathbb{C}$ is $2\pi-$periodic and analytic on $\mathbb{R}$ then

$$\bigl|I(f) - T_N(f)\bigr| \;\le\; \frac{4\pi c}{\mathrm{e}^{\sigma N} - 1}\,, \tag{1.78b}$$

where $\sigma > 0$ denotes the width of the strip $\mathbb{R} + i(-\sigma/2, \sigma/2) \subset \mathbb{C}$ into which f can be extended analytically and c is a bound on $|f|$ in this strip.

For a proof we refer to Kress [76].

Second, we consider an arbitrary open line source which leads to an integral of the form (1.74) where f is not necessarily periodic. The most common quadrature rule is the (composite) **Simpson's rule**:

Theorem 1.18. *Let $n \in \mathbb{N}$ be even and $t_j = a + jh$, $j = 0, \ldots, n$, where $h = \frac{b-a}{n}$, and*

$$S_n(f) \;:=\; \frac{h}{3}\bigl[f(t_0) + 4f(t_1) + 2f(t_2) + \cdots + 2f(t_{n-2}) + 4f(t_{n-1}) + f(t_n)\bigr]\,. \tag{1.79}$$

For $f \in C^4[a,b]$ one has the error estimate

$$\bigl|I(f) - S_n(f)\bigr| \;\le\; \frac{b-a}{180}\, h^4\, \bigl\|f^{(4)}\bigr\|_{C[a,b]}\,. \tag{1.80}$$

For a proof we refer again to Kress [76].

It is important to note that the order 4 does not increase for smoother functions f. We describe two possibilities for quadrature formulas which automatically adopt the order of convergence to the smoothness of f.

The first is the **Gauss-Legendre method**. By the change of variables $s = \frac{t}{2}(b-a) + \frac{1}{2}(b+a)$, $t \in [-1,+1]$, i.e.

$$\int_a^b f(s)\,ds \;=\; \frac{b-a}{2}\int_{-1}^{+1} f\bigl(t(b-a)/2 + (b+a)/2\bigr)\,dt\,,$$

we can assume that $[a,b] = [-1,+1]$. The **Legendre polynomials** are recursively defined by $P_0(t) := 1$, $P_1(t) := t$, $t \in \mathbb{R}$, and

$$P_{n+1}(t) \;:=\; \frac{2n+1}{n+1}\,t\,P_n(t) \;-\; \frac{n}{n+1}\,P_{n-1}(t)\,, \quad t \in \mathbb{R}\,,\ n = 1,2,\ldots.$$

They are orthogonal in $L^2(-1,+1)$. From the theory of orthogonal polynomials it can be shown that the zeros $t_1^{(n)},\ldots,t_n^{(n)}$ of P_n are all simple, real, and lie in the interval $(-1,+1)$. There is, however, no analytical expression for these $t_j^{(n)}$. The weights $w_j^{(n)}$ are determined by the requirement that

$$\int_{-1}^{+1} P_k(t)\,dt \;=\; \sum_{j=1}^{n} w_j^{(n)}\,P_k(t_j^{(n)}) \quad \text{for all } k = 0,\ldots,n-1\,.$$

It can be shown that this system is uniquely solvable and the weights are all positive. For $n = 1$ and $n = 2$ the nodes and weights are given by $t_1^{(1)} = 0$, $w_1^{(1)} = 2$, and $t_1^{(2)} = -1/\sqrt{3}$, $t_2^{(2)} = +1/\sqrt{3}$, $w_1^{(2)} = w_2^{(2)} = 1$, respectively. For higher values of n we refer to Abramowitz, Stegun [1] for approximate values of $t_j^{(n)}$ and $w_j^{(n)}$. Then there exists a convergence result analogously to Theorem 1.17. In practice, however, one often prefers composite Gaussian rules to avoid the computation of the nodes and weights. We refer to Kress [76] for more details.

In a second class of methods for the numerical integration of general, nonperiodic, functions one transforms the integral into one with periodic integrand. Let the function $w : [0,2\pi] \longrightarrow [a,b]$ be bijective[5], strictly monotonically increasing and infinitely often differentiable. Then we substitute $s = w(t)$ in (1.74) and have

$$\int_a^b f(s)\,ds \;=\; \int_0^{2\pi} g(t)\,dt \quad \text{with} \quad g(t) = w'(t)\,f\bigl(w(t)\bigr)\,,\ 0 \le t \le 2\pi\,. \qquad (1.81)$$

Now assume that w has derivatives

[5] that is, one-t-one and onto

$$w^{(j)}(0) \;=\; w^{(j)}(2\pi) \;=\; 0\,, \quad j = 1,\ldots,p-1\,,$$

for some odd $p \geq 3$. If $f \in C^p[a,b]$ then g is 2π–periodic and p–times continuously differentiable with

$$g^{(j)}(0) \;=\; g^{(j)}(2\pi) \;=\; 0\,, \quad j = 1,\ldots,p-2\,.$$

Therefore, we can apply the trapezoidal rule to $\int_0^{2\pi} g(t)\,dt$, i.e.

$$\int\limits_a^b f(s)\,ds \;\approx\; Q_N(f) \;:=\; \frac{\pi}{N} \sum_{j=0}^{2N+1} w'(t_j)\, f\big(w(t_j)\big)\,, \tag{1.82}$$

where $t_j = j\frac{\pi}{N}$, $j = 0,\ldots,2N-1$. This quadrature formula has nodes $w(t_j)$ and weights $\frac{\pi}{N}\, w'(t_j)$. Application of Theorem 1.17 yields the error estimate

$$\big|I(f) - Q_N(f)\big| \;\leq\; \frac{c}{N^{p-2}} \int\limits_0^{2\pi} \left| \frac{d^{p-2}}{dt^{p-2}}\,(f \circ w)(t) \right| dt\,. \tag{1.83}$$

There exist many examples for substitutions $s = w(t)$. In the following numerical experiments we have chosen

$$w_p(t) \;=\; a \;+\; \frac{t^p}{t^p + (2\pi - t)^p}\,(b-a)\,, \quad t \in [0, 2\pi]\,,$$

for $p = 3, 5, 7$.

Example 1.19. We illustrate the convergence properties for the linear line factor

$$\int\limits_{-1}^{+1} \psi(s)\,\mathrm{e}^{-ikst}\,ds \quad \text{with} \quad \psi(s) \;:=\; i\sin(\pi s)\,,\ |s| \leq 1\,,$$

where the exact solution is known to be

$$\int\limits_{-1}^{+1} \psi(s)\,\mathrm{e}^{-ikst}\,ds \;=\; \frac{\sin(\pi - kt)}{\pi - kt} \;-\; \frac{\sin(\pi + kt)}{\pi + kt}\,.$$

In the following tables we list the errors for some values of p and kt:

	$p\;=\;3$			
N	$kt = 3$	$kt = 6$	$kt = 9$	$kt = 12$
2	$9.7691942e-01$	$1.4727376e-01$	$1.4382097e-01$	$1.4642124e-01$
4	$2.1311168e-01$	$6.6557897e-01$	$2.9781892e-01$	$7.4794306e-01$
8	$2.0013140e-04$	$5.5985148e-03$	$5.7711962e-02$	$2.6596132e-01$
16	$3.9574788e-11$	$3.6551596e-10$	$4.4367871e-08$	$1.9560116e-06$
32	$6.0995653e-13$	$1.2080892e-12$	$1.7816998e-12$	$2.3191483e-12$
64	$9.3258734e-15$	$1.9012569e-14$	$2.7901292e-14$	$3.6137759e-14$

	$p = 5$			
N	$kt = 3$	$kt = 6$	$kt = 9$	$kt = 12$
2	$1.0191761e + 00$	$6.8101247e - 02$	$3.7743795e - 02$	$2.6863829e - 02$
4	$8.4946538e - 01$	$3.5296713e - 01$	$3.4617122e - 01$	$2.5963205e - 01$
8	$7.3449332e - 02$	$3.8739395e - 01$	$6.2345110e - 01$	$1.8167665e - 01$
16	$1.0600692e - 05$	$4.3842412e - 04$	$7.0002669e - 03$	$5.4326473e - 02$
32	$1.5543122e - 15$	$8.4617036e - 13$	$1.2304045e - 10$	$8.0069052e - 09$
64	$2.2204460e - 16$	$6.9388939e - 17$	$4.0939474e - 16$	$3.4694470e - 16$

	$p = 7$			
N	$kt = 3$	$kt = 6$	$kt = 9$	$kt = 12$
2	$1.0196326e + 00$	$6.7200807e - 02$	$3.6424352e - 02$	$2.5161877e - 02$
4	$9.9941155e - 01$	$1.0577645e - 01$	$8.9793736e - 02$	$8.8398572e - 02$
8	$4.0282892e - 01$	$7.1861750e - 01$	$1.1515977e - 01$	$5.2077521e - 01$
16	$2.5429396e - 03$	$3.6830896e - 02$	$2.1031953e - 01$	$5.1202900e - 01$
32	$1.4103223e - 09$	$1.7583851e - 07$	$7.9043821e - 06$	$1.7263055e - 04$
64	$4.4408921e - 16$	$4.3021142e - 16$	$7.7021722e - 16$	$1.0096091e - 15$

We observe that high values of p lead to worse results, in particular for large values of kt. This seems to contradict the theory but is a result of cancellations due to the oscillations of the integrand.

For comparisons, we have also computed this example with Simpson's rule. The result is given in the following table (here $n = 2N$):

n	$kt = 3$	$kt = 6$	$kt = 9$	$kt = 12$
4	$3.1035363e - 01$	$1.2097315e - 01$	$1.2669697e + 00$	$3.9768914e - 01$
8	$9.7345528e - 04$	$1.0735333e - 02$	$2.8950046e - 01$	$1.4252548e - 01$
16	$4.7656766e - 05$	$3.3266175e - 04$	$1.2416022e - 03$	$3.6517059e - 03$
32	$2.8206792e - 06$	$1.8268067e - 05$	$5.9955240e - 05$	$1.4367000e - 04$
64	$1.7396911e - 07$	$1.1074264e - 06$	$3.5354250e - 06$	$8.1439882e - 06$
128	$1.0837294e - 08$	$6.8695078e - 08$	$2.1784657e - 07$	$4.9715128e - 07$

One clearly observes the superiority of the transformation rule.

1.6 Conclusion

For the case of arrays, there are a number of parameters which describe the antenna. Among are, (1) the number of antenna elements which comprise the array; (2) the geometric configuration of the array; (3) the particular location of the elements; and (4) the excitation of the elements of the array. For the case of line sources, the parameters are the shape of the curve and the current distribution. Likewise, the description of the behavior of the array of line source is specified by a number of parameters (1) the radiation pattern; (2) the directivity; (3) the power gain; (4) the impedance; (5) the side lobe levels; and (6) the beam-width.

We have illustrated in this chapter some of the common optimization problems which have been discussed with respect to arrays of dipoles and to line sources. The exposition has not be rigorous, in so far as we have not carefully derived the various expressions for the electromagnetic quantities that we considered. Nevertheless, we have indicated some of the basic themes that will concern us in the following chapters.

Our object, in the remainder of this work, is to present a more rigorous approach to the general problem of antenna optimization which includes a variety of physical configurations. What should be clear from the preceding pages is that the *problems of antenna optimization can be put in a general framework.* Specifically, we can view the antenna optimization problem as a set of admissible inputs of which can be fed into a system consisting of an "antenna" which then produces a state (the far field) and which can be associated with the feedings by means of an operator

$$\mathcal{K} : \boldsymbol{a} \mapsto f \quad \text{or} \quad \mathcal{K} : \psi \mapsto f$$

from some space X of inputs into the space of factors which is usually the space $C(S^{d-1})$, $d = 2$ or 3, of continuous functions on the unit sphere. If the factors are independent of ϕ as in the case of a linear array or linear line source along the $z-$axis then we can as well take the output space to be $C[0, \pi]$ or $C[-1, +1]$ by identifying $\theta \in [0, \pi]$ with $t = \cos\theta \in [-1, +1]$.

In the next chapters we will make a systematic presentation of Maxwell's equations and their use in order to describe radiated fields from antenna sources. We will then investigate various problems of optimization, both theoretically and numerically, which have relevance for practical design and, as well, will raise interesting mathematical questions.

2 Discussion of Maxwell's Equations

2.1 Introduction

The history of electromagnetism, and in particular that part which lead in the 19^{th} century to the formulation of the equations governing the electromagnetic fields, is studded with the names of the leading scientists of the time. Starting with Gauss, there was a more or less ongoing effort to understand the relationships and to model their interactions, and included efforts by B. Riemann, W. Thompson, M. Faraday, as well as Neumann, Kirchhoff and Weber and Helmholtz. Building on all this work, Maxwell's great insight was the introduction of the notion of the so-called displacement current, $\boldsymbol{D}$, a generalization of Faraday's idea of charge polarization or displacement. Using Faraday's work and the ideas of elastic continua, Maxwell developed his famous equations, and noting the close agreement between the electric ratio c and the velocity of light, asserted the coincidence of the two phenomena.[1]

His, and his contemporaries' assumption was that it was necessary to postulate the existence of some medium (the ether) through which the electromagnetic waves would be propagated, and idea finally put to rest by the famous experiment of Michaelson and Merely.

In this chapter, we will present Maxwell's equations and some of the related theory of electromagnetic potentials.

2.2 Geometry of the Radiating Structure

We will consider a prescribed radiating structure S as some subset of the usual three-dimensional Euclidean space $\mathbb{R}^3$ which represents a physical body

[1] It is interesting to realize that much of the development of the theory by Thompson, Maxwell, and others, which culminated in the model described by what are now called Maxwell's Equations, was made with the fluid dynamical model firmly in mind. We refer the interested reader to the interesting historical essay on the work leading to this system of equations in [143].

capable of supporting a flow of electric current. The following shapes are of particular interest:

(a) S consists of finitely many points with position vectors $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m \in \mathbb{R}^3$. Such a configuration will be called an *antenna array.* The points could lie on a line (linear array) or in a plane (plane array) or be more generally distributed within a three dimensional volume.
(b) S is a curve of finite length in $\mathbb{R}^3$ (a wire) or a collection of such curves. This case could also be considered as the limiting case of (a) where the number m of points tends to infinity and the distances between them tend to zero.
(c) S is a connected part of the boundary ∂D of some open and bounded subset D of $\mathbb{R}^3$. This class includes, as particular cases, both *reflector-* and *slot antennas* and, more generally, so called *conformal antennas.*
(d) S is an infinite cylinder with axis in some direction (e.g. in the x_3-direction) with constant cross section S' which could be considered as a subset of the two dimensional Euclidean space $\mathbb{R}^2$. S' could be a disc, an annulus, a curve, or even a more complicated domain.

The first chapter was devoted to an elementary discussion of examples which fall under the first category. In this chapter, we will give a discussion of the equations governing electromagnetic radiation from structures of a general type which will include all the cases enumerated above.

2.3 Maxwell's Equations in Integral Form

Electromagnetic wave propagation is described by particular equations relating five vector fields $\boldsymbol{\mathcal{E}}$, $\boldsymbol{\mathcal{D}}$, $\boldsymbol{\mathcal{H}}$, $\boldsymbol{\mathcal{B}}$, $\boldsymbol{\mathcal{J}}$ and the scalar field ρ, where $\boldsymbol{\mathcal{E}}$ and $\boldsymbol{\mathcal{D}}$ denote the **electric field** (in V/m) and **electric induction** (in As/m^2) respectively, while $\boldsymbol{\mathcal{H}}$ and $\boldsymbol{\mathcal{B}}$ denote the **magnetic field** (in A/m) and **magnetic induction** (in $Vs/m^2 = T$ =Tesla). Likewise, $\boldsymbol{\mathcal{J}}$ and ρ denote the **current** (in A/m^2) and **charge distribution** (in As/m^3) of the medium. Here and throughout the book we use the **rationalized MKS-system**, i.e. V, A, m and s (see [130], section 1.8). All fields will be assumed to depend both on the space variable $\boldsymbol{x} \in \mathbb{R}^3$ and on the time variable $t \in \mathbb{R}$.

The actual equations that govern the behavior of the electromagnetic field, first completely formulated by Maxwell, may be expressed easily in integral form. Such a formulation, which has the advantage of being closely connected to the physical situation, has been used to effectively by a number of authors, in particular by Sommerfeld [125] and by Müller [105]. The more familiar differential form of Maxwell's equations can be derived very easily from the integral relations as we will see below in Section 2.5.

In order to write these integral relations, we begin by letting S be a connected smooth surface with boundary ∂S in the interior of a region D where electromagnetic waves propagate. In particular, we require that the unit normal vector $\boldsymbol{n}(\boldsymbol{x})$ for $\boldsymbol{x} \in S$ be continuous and directed always into "one side" of S, which we call the positive side of S. By $\boldsymbol{t}(\boldsymbol{x})$ we denote the unit vector tangent to the boundary of S at $\boldsymbol{x} \in \partial S$. This vector, lying in the tangent plane of S together with a vector $\boldsymbol{\nu}(\boldsymbol{x})$, $\boldsymbol{x} \in \partial S$, normal to ∂S is oriented so as to form a mathematically positive system (i.e. $\boldsymbol{t}$ is directed counterclockwise when we sit on the positive side of S). Furthermore, let $\Omega \in \mathbb{R}^3$ be an open set with boundary $\partial\Omega$ and outer unit normal vector $\boldsymbol{n}(\boldsymbol{x})$ at $\boldsymbol{x} \in \partial\Omega$. Then Maxwell's equations in integral form state:

$$\int\limits_{\partial S} \boldsymbol{\mathcal{H}} \cdot \boldsymbol{t}\, d\ell = \frac{\partial}{\partial t} \int\limits_{S} \boldsymbol{\mathcal{D}} \cdot \boldsymbol{n}\, dS + \int\limits_{S} \boldsymbol{\mathcal{J}} \cdot \boldsymbol{n}\, dS \quad \text{(Ampère's Law)} \tag{2.1a}$$

$$\int\limits_{\partial S} \boldsymbol{\mathcal{E}} \cdot \boldsymbol{t}\, d\ell = -\frac{\partial}{\partial t} \int\limits_{S} \boldsymbol{\mathcal{B}} \cdot \boldsymbol{n}\, dS \quad \text{(Law of Induction)} \tag{2.1b}$$

$$\int\limits_{\partial\Omega} \boldsymbol{\mathcal{D}} \cdot \boldsymbol{n}\, dS = \int\limits_{\Omega} \rho\, dV \qquad \text{(Gauss' Electric Law)} \tag{2.1c}$$

$$\int\limits_{\partial\Omega} \boldsymbol{\mathcal{B}} \cdot \boldsymbol{n}\, dS = 0 \quad \text{(Gauss' Magnetic Law)} \tag{2.1d}$$

The initial goal of using such equations to model the electromagnetic field is to enable us to determine uniquely the five field quantities which result from a given distribution of currents and charges. From this point of view, the four equations are incomplete and must be supplemented by equations which describe the interaction between the fields and the medium through which the fields propagate. These **constitutive relations**, characteristic of the medium, may be either linear or nonlinear. In this book we will deal exclusively with linear constitutive relations which we describe in the next section.

2.4 The Constitutive Relations

In light of the preceding comments, we will consider electromagnetic wave propagation in **linear, isotropic** media. This means, first, that there exist linear relationships (the **constitutive relations**) between $\boldsymbol{\mathcal{E}}$ and $\boldsymbol{\mathcal{D}}$, $\boldsymbol{\mathcal{H}}$, and $\boldsymbol{\mathcal{B}}$:

$$\boldsymbol{\mathcal{D}} \;=\; \epsilon\, \boldsymbol{\mathcal{E}}\,, \quad \text{and} \quad \boldsymbol{\mathcal{B}} \;=\; \mu\, \boldsymbol{\mathcal{H}}\,. \tag{2.2}$$

In general, the quantities ϵ and μ may be space dependent, but we assume that they are independent of time and of direction and are therefore scalar (as opposed to tensor) quantities. Hence the term *isotropic.*

The **permittivity** or **dielectric constant**, ϵ, has a unit As/Vm, and is related to the ability of the medium to sustain an electric charge. Its value, ϵ_0, in a vacuum has been experimentally determined and is approximately

$8.854 \cdot 10^{-12}\,As/Vm$ while that, say, in fused quartz it is approximately $3.545 \cdot 10^{-11}\,As/Vm$.

The **magnetic permeability** for most substances, μ, is close to its value in vacuo $\mu_0 = 4\pi \cdot 10^{-7}\,Vs/Am$. Those substances for which μ is significantly different from this value are called magnetic, either **paramagnetic** or **diamagnetic** if $\mu > \mu_0$ or $\mu < \mu_0$, respectively. In the following, however, we always will assume that $\mu = \mu_0$.

Usually ϵ and μ are independent of the field strength although in some important situations this is not the case. As we will mention below, one concomitant effect of attempting to synthesize a highly focused beam, is the storage of power close to the antenna itself, which may degrade performance because of dramatic alterations in these constitutive parameters of the atmosphere.

The quantity $c_0 := 1/\sqrt{\epsilon_0\mu_0}$ has the dimension of velocity. It is a consequence of the field equations that this quantity is the velocity of propagation of the electromagnetic field disturbance through free space. Experimental measurements have shown that, in vacuo, this velocity is the same as that of light and hence $c_0 \approx 2.9979 \cdot 10^8\,m/s$.

Two special cases will be considered in the following: media in which the constitutive parameters vary smoothly, and media in which there are manifolds of discontinuity (interfaces) of these parameters. In a medium where ϵ and μ vary smoothly, Maxwell's equations are equivalent to a system of partial differential equations. In the second case where an interface exists, the behaviour of the constitutive parameters together with the correct choice of S and Ω lead to boundary conditions for these equations.

2.5 Maxwell's Equations in Differential Form

First, we consider a region D where μ and ϵ are constant (*homogeneous medium*) or at least continuous. In regions where the vector fields are smooth functions we can apply the Stokes and Gauss theorems for surfaces S and solids Ω lying completely in D:

$$\int_S \operatorname{curl} \boldsymbol{F} \cdot \boldsymbol{n}\, dS = \int_{\partial S} \boldsymbol{F} \cdot \boldsymbol{t}\, d\ell \quad \text{(Stokes)}, \tag{2.3}$$

$$\int_\Omega \operatorname{div} \boldsymbol{F}\, dV = \int_{\partial\Omega} \boldsymbol{F} \cdot \boldsymbol{n}\, dS \quad \text{(Gauss)}, \tag{2.4}$$

where $\boldsymbol{F}$ denotes one of the fields $\boldsymbol{\mathcal{H}}$, $\boldsymbol{\mathcal{E}}$, $\boldsymbol{\mathcal{B}}$ or $\boldsymbol{\mathcal{D}}$. With these formulas we can eliminate the boundary integrals in (2.1a-2.1d). We then use the fact that we can vary the surface S and the solid Ω in D arbitrarily. By equating the integrands we are led to Maxwell's equations in **differential form** so that Ampère's Law, the Law of Induction and Gauss' Electric and Magnetic Laws, respectively, become:

$$\operatorname{curl} \boldsymbol{\mathcal{H}} = \frac{\partial}{\partial t}\boldsymbol{\mathcal{D}} + \boldsymbol{\mathcal{J}} \tag{2.5a}$$

$$\operatorname{curl} \boldsymbol{\mathcal{E}} = -\frac{\partial}{\partial t}\boldsymbol{\mathcal{B}} \tag{2.5b}$$

$$\operatorname{div} \boldsymbol{\mathcal{D}} = \rho \tag{2.5c}$$

$$\operatorname{div} \boldsymbol{\mathcal{B}} = \boldsymbol{o} \tag{2.5d}$$

Taking the divergence of (2.5a), using (2.5c), and noting that div curl $= 0$ we derive an equation relating the current and charge densities:

$$\operatorname{div} \boldsymbol{\mathcal{J}} + \frac{\partial}{\partial t}\rho = 0. \tag{2.6}$$

We may consider (2.6), as analogous to the continuity or conservation equation in fluid dynamics. It expresses the fact that charge is conserved in the neighborhood of any point.

The current density $\boldsymbol{\mathcal{J}}$ commonly consists of two terms: one, $\boldsymbol{\mathcal{J}}_e$, associated with external sources of electromagnetic disturbances and the other, $\boldsymbol{\mathcal{J}}_c$, associated with conduction currents produced as a result of the electric field. In many cases we will be considering source free regions for which $\boldsymbol{\mathcal{J}}_e = \boldsymbol{o}$.

To the linear constitutive relations, we add a third, namely **Ohm's Law**, which relates the quantities $\boldsymbol{\mathcal{J}}_c$ and $\boldsymbol{\mathcal{E}}$ by a linear relation,

$$\boldsymbol{\mathcal{J}}_c = \sigma \boldsymbol{\mathcal{E}}. \tag{2.7}$$

The scalar function σ which is called the **conductivity** has units of $\frac{A}{Vm}$. Substances for which σ is not negligibly small are called **conductors**. Metals, for example, are good conductors as is brine. In general, the conductivity in metals decreases with increasing temperature, but in the case of other materials, the **semiconductors**, conductivity increases with temperature over a wide range.

By way of contrast, substances for which σ is negligibly small are called **dielectrics** or **insulators**. For such substances, their electromagnetic properties are completely determined by the other constitutive parameters ϵ and μ. For the purposes of analysis, it is often convenient to approximate good conductors by **perfect conductors**, characterized by $\sigma = \infty$, and good dielectrics by **perfect dielectrics** characterized by $\sigma = 0$. Examples of conductors are given in the following table:

The constitutive relations (2.2) and Ohm's law (2.7) allow us to eliminate $\boldsymbol{\mathcal{D}}$, $\boldsymbol{\mathcal{B}}$ and $\boldsymbol{\mathcal{J}}_c$ from Maxwell's equations. Thus in a linear, isotropic, conducting medium we see that the propagation of the electromagnetic field is described by

Material	**Conductivity** in siemens/meter at 20°C
Copper, annealed	$5.8005 \cdot 10^7$
Gold	$4.10 \cdot 10^7$
Steel	$0.5 - 1.0 \cdot 10^7$
Nickel	$1.28 \cdot 10^7$
Silver	$6.139 \cdot 10^7$
Tin	$0.869 \cdot 10^7$
Glass, ordinary	10^{-12}
Mica	$10^{-11} - 10^{-15}$
Porcelain	$3 \cdot 10^{-13}$
Quartz, fused	$< 2 \cdot 10^{-17}$
Methyl Alcohol	$7.1 \cdot 10^{-4}$
Water, distilled (18°C)	$2 \cdot 10^{-4}$
Sea Water	$3 - 5$

Table 2.1. Table of Conductivities

$$\operatorname{curl}\mathcal{H} = \epsilon\frac{\partial}{\partial t}\mathcal{E} + \sigma\mathcal{E} + \mathcal{J}_e\,, \tag{2.8a}$$

$$\operatorname{curl}\mathcal{E} = -\mu\frac{\partial}{\partial t}\mathcal{H}\,, \tag{2.8b}$$

$$\operatorname{div}(\epsilon\mathcal{E}) = \rho\,, \tag{2.8c}$$

$$\operatorname{div}(\mu\mathcal{H}) = 0\,. \tag{2.8d}$$

Another remarkable equation which holds in isotropic *homogeneous* conductors in source free regions follows directly from these equations. Indeed observing that in this situation $\mathcal{J}_e = \boldsymbol{o}$, taking the divergence of the equation (2.8a), and differentiating (2.8c) with respect to time, we find the two equations

$$0 = \epsilon\operatorname{div}\frac{\partial\mathcal{E}}{\partial t} + \sigma\operatorname{div}\mathcal{E}$$

and

$$\epsilon \operatorname{div} \frac{\partial \mathcal{E}}{\partial t} = \frac{\partial \rho}{\partial t}.$$

In this manner we arrive at the differential equation

$$\frac{\sigma}{\epsilon}\rho + \frac{\partial}{\partial t}\rho = 0,$$

which (since σ and ϵ are constant) has the unique solution for $t > 0$:

$$\rho(x,t) = \rho(x,0)\,\mathrm{e}^{-\sigma t/\epsilon}.$$

We can interpret σ/ϵ as an angular frequency characteristic of the medium, and call the reciprocal ϵ/σ the **relaxation time**. For copper this is approximately $1.5 \cdot 10^{-19} s$. Thus, for an electric disturbance incident from the exterior of a conductor, the electric charge density falls off exponentially with time. From this analysis it is clear that a metallic conductor does not support a charge and it is a reasonable approximation to replace (2.8c) with $\operatorname{div}(\epsilon\mathcal{E}) = 0$.

2.6 Energy Flow and the Poynting Vector

The description of the performance of antennas, which is the central theme of this book, often involves numerical measures which depend for their definition on the notion of power contained in the field. The power in the electromagnetic field is most often described using the **Poynting vector** $\mathcal{S} = \mathcal{E} \times \mathcal{H}$, and we are interested next in understanding how it arises. From the vector identity

$$\operatorname{div}(\mathcal{E} \times \mathcal{H}) = \mathcal{H} \cdot \operatorname{curl} \mathcal{E} - \mathcal{E} \cdot \operatorname{curl} \mathcal{H}$$

and Maxwell's equations (2.5a), (2.5b) we get immediately

$$\operatorname{div}(\mathcal{E} \times \mathcal{H}) = -\mathcal{H} \cdot \frac{\partial \mathcal{B}}{\partial t} - \mathcal{E} \cdot \frac{\partial \mathcal{D}}{\partial t} - \mathcal{E} \cdot \mathcal{J}, \tag{2.9}$$

which is valid in any medium. In light of the constitutive relations, the terms involving time derivatives in (2.9) lead to

$$\begin{aligned}\mathcal{H} \cdot \frac{\partial \mathcal{B}}{\partial t} + \mathcal{E} \cdot \frac{\partial \mathcal{D}}{\partial t} &= \mu \mathcal{H} \cdot \frac{\partial \mathcal{H}}{\partial t} + \epsilon \mathcal{E} \cdot \frac{\partial \mathcal{E}}{\partial t} \\ &= \mu \frac{1}{2} \frac{\partial}{\partial t}(\mathcal{H} \cdot \mathcal{H}) + \epsilon \frac{1}{2} \frac{\partial}{\partial t}(\mathcal{E} \cdot \mathcal{E}) \\ &= \frac{1}{2} \frac{\partial}{\partial t}(\mathcal{B} \cdot \mathcal{H} + \mathcal{D} \cdot \mathcal{E}).\end{aligned}$$

It can be shown (see e.g., [86]) that this equation expresses conservation of energy. In particular the right hand side of this equation represents the rates of increase of electric and magnetic internal energies $U_e := \frac{1}{2}\mathcal{D} \cdot \mathcal{E}$ and $U_m :=$

$\frac{1}{2}\mathcal{H}\cdot\mathcal{B}$, respectively, per unit volume. Using the Poynting vector $\boldsymbol{\mathcal{S}} = \mathcal{E}\times\mathcal{H}$ we can rewrite (2.9) as

$$\frac{\partial}{\partial t}(U_e + U_m) \;+\; \operatorname{div}\boldsymbol{\mathcal{S}} \;=\; -\mathcal{E}\cdot\mathcal{J}\,. \tag{2.10}$$

The term $\mathcal{E}\cdot\mathcal{J}$ is the rate per unit volume at which the electric field is doing work. In the absence of external currents this will represent heat dissipation. We note that this equation takes the form of a conservation law.

Using Gauss' Theorem (2.4) in (2.10), we find that, for any volume Ω with smooth surface $\partial\Omega$,

$$\frac{\partial}{\partial t}\int\limits_{\Omega}(U_e + U_m)\,dV \;+\; \int\limits_{\Omega}\mathcal{J}\cdot\mathcal{E}\,dV \;+\; \int\limits_{\partial\Omega}\boldsymbol{\mathcal{S}}\cdot\boldsymbol{n}\,dS \;=\; 0\,. \tag{2.11}$$

This equation is sometimes called **Poynting's Theorem** or the energy balance equation. Setting $W := \int_\Omega (U_e + U_m)dV$ and $Q = \int_\Omega \mathcal{J}\cdot\mathcal{E}\,dV$, then W and Q represent, respectively, the total energy and the resistive dissipation of energy, called **Joule heat** in the conductor. There is a further decrease of energy if the field extends to the bounding surface of the volume, so the surface integral

$$\mathcal{P} \;=\; \int\limits_{\partial\Omega}\boldsymbol{\mathcal{S}}\cdot\boldsymbol{n}\,dS \;=\; \int\limits_{\partial\Omega}(\mathcal{E}\times\mathcal{H})\cdot\boldsymbol{n}\,dS \tag{2.12}$$

in (2.11) must represent the flow of energy across the boundary. Therefore, we may think of $\boldsymbol{\mathcal{S}} = \mathcal{E}\times\mathcal{H}$ as representing the amount of energy crossing the boundary per second, per unit area.

It is important to understand, however, that the vector $\boldsymbol{\mathcal{S}}$ is a *construct*; the actual quantity of importance in the energy balance equation is $\boldsymbol{\mathcal{S}}\cdot\boldsymbol{n}$. In light of Gauss' theorem, we can add the curl of an arbitrary field without changing the value of the integral in (2.11) and so the choice of the vector $\boldsymbol{\mathcal{S}}$ is not unique.

From this analysis, one may conclude that the conductivity of a medium is connected to the appearance of Joule heat. Thermodynamically irreversible, this process transforms the electromagnetic energy into heat and, consequently, the wave is attenuated as it penetrates the conductor. This effect is particularly pronounced in metals with high conductivity. This leads to the so-called **skin effect** which we will make more precise below in our discussion of time-harmonic fields.

2.7 Time Harmonic Fields

From now on we assume that all fields vary periodically in time with the same angular frequency $\omega = 2\pi/T$ and period T. This could be insured by assuming periodic time dependence of the applied external currents or fields.

It is very convenient to use the complex representation of the fields in the form

$$\begin{aligned} \mathcal{E}(\boldsymbol{x},t) &= \operatorname{Re}\left(\boldsymbol{E}(\boldsymbol{x})\,\mathrm{e}^{-i\omega t}\right), & \mathcal{D}(\boldsymbol{x},t) &= \operatorname{Re}\left(\boldsymbol{D}(\boldsymbol{x})\,\mathrm{e}^{-i\omega t}\right), \\ \mathcal{H}(\boldsymbol{x},t) &= \operatorname{Re}\left(\boldsymbol{H}(\boldsymbol{x})\,\mathrm{e}^{-i\omega t}\right), & \mathcal{B}(\boldsymbol{x},t) &= \operatorname{Re}\left(\boldsymbol{B}(\boldsymbol{x})\,\mathrm{e}^{-i\omega t}\right), \\ \mathcal{J}(\boldsymbol{x},t) &= \operatorname{Re}\left(\boldsymbol{J}(\boldsymbol{x})\,\mathrm{e}^{-i\omega t}\right), \end{aligned}$$

as well as for ρ. Here, $\boldsymbol{E}$, $\boldsymbol{D}$, $\boldsymbol{H}$, $\boldsymbol{B}$ and $\boldsymbol{J}$ are now space dependent complex vector fields. By using these formulas the derivative with respect to time transforms into multiplication by $-i\omega$. Thus, Maxwell's equations (2.8a)–(2.8d) in conducting and isotropic media read for the space dependent parts

$$\operatorname{curl}\boldsymbol{H} = (-i\omega\epsilon + \sigma)\,\boldsymbol{E} + \boldsymbol{J}_e\,, \tag{2.13a}$$

$$\operatorname{curl}\boldsymbol{E} = i\omega\mu\,\boldsymbol{H}\,, \tag{2.13b}$$

$$\operatorname{div}(\epsilon\boldsymbol{E}) = \rho\,, \tag{2.13c}$$

$$\operatorname{div}\boldsymbol{H} = 0\,. \tag{2.13d}$$

We assume, for the following analysis, that $\mu = \mu_0$ is the magnetic permeability of vacuo. We remark that in this case (2.13d) follows directly from (2.13b) while for homogeneous media equation (2.13c) follows from (2.13a) (with $\rho = \operatorname{div}\boldsymbol{J}_e/(i\omega - \sigma/\epsilon)$). In these cases, both of them can be omitted from the system. In source free media in particular, $\operatorname{div}\boldsymbol{E} = 0$ and therefore no distributed charge ρ exists.

By taking the curl again we can eliminate either $\boldsymbol{E}$ or $\boldsymbol{H}$ from the system:

$$-\operatorname{curl}^2\boldsymbol{E} + i\omega\,\mu_0\,(\sigma - i\omega\,\epsilon)\,\boldsymbol{E} = -i\omega\mu_0\,\boldsymbol{J}_e\,, \quad \operatorname{div}(\epsilon\boldsymbol{E}) = \rho\,, \tag{2.14a}$$

$$-\operatorname{curl}\left(\frac{1}{\sigma - i\omega\epsilon}\operatorname{curl}\boldsymbol{H}\right) + i\omega\mu_0\,\boldsymbol{H} = -\operatorname{curl}\left(\frac{1}{\sigma - i\omega\epsilon}\,\boldsymbol{J}_e\right). \tag{2.14b}$$

With $\boldsymbol{E}$ or $\boldsymbol{H}$ from (2.14a) or (2.14b), respectively, one has to compute $\boldsymbol{H}$ or $\boldsymbol{E}$ by formulas (2.13b) or (2.13a), respectively.

It is convenient to introduce the complex **wave number** $k \in \mathbb{C}$ by

$$k^2 = i\omega\,\mu_0\,(\sigma - i\omega\,\epsilon) = \omega^2\mu_0\,\epsilon + i\omega\,\mu_0\,\sigma\,. \tag{2.15}$$

Since only k^2 occurs we can choose that branch of the square root with $\operatorname{Re} k \geq 0$ and also $\operatorname{Im} k \geq 0$.

In *homogeneous* media ϵ and σ are constant. In this case we note that $\operatorname{curl}^2 = \operatorname{grad}\operatorname{div} - \Delta$ and arrive at the inhomogeneous (vector-) **Helmholtz equations**

$$\Delta\boldsymbol{E} + k^2\,\boldsymbol{E} = \nabla\rho/\epsilon - i\omega\mu_0\,\boldsymbol{J}_e\,, \quad \Delta\boldsymbol{H} + k^2\,\boldsymbol{H} = -\operatorname{curl}\boldsymbol{J}_e\,, \tag{2.16}$$

$\operatorname{div}\boldsymbol{E} = \rho/\epsilon$, and $\operatorname{div}\boldsymbol{H} = 0$.

Writing $\boldsymbol{\mathcal{E}}(\boldsymbol{x},t) = \frac{1}{2}[\boldsymbol{E}(\boldsymbol{x})\exp(-i\omega t) + \overline{\boldsymbol{E}(\boldsymbol{x})}\exp(i\omega t)]$ and analogously for $\boldsymbol{\mathcal{H}}$ we have for the Poynting vector $\boldsymbol{\mathcal{S}}$ from Section 2.7 (after a short calculation)

$$\begin{aligned}\boldsymbol{\mathcal{S}}(\boldsymbol{x},t) &= \boldsymbol{\mathcal{E}}(\boldsymbol{x},t) \times \boldsymbol{\mathcal{H}}(\boldsymbol{x},t) \\ &= \frac{1}{2}\operatorname{Re}[\boldsymbol{E}(\boldsymbol{x}) \times \overline{\boldsymbol{H}(\boldsymbol{x})}] + \frac{1}{2}\operatorname{Re}\left[\boldsymbol{E}(\boldsymbol{x}) \times \boldsymbol{H}(\boldsymbol{x})\, \mathrm{e}^{-2i\omega t}\right].\end{aligned}$$

The first term $\frac{1}{2}\operatorname{Re}[\boldsymbol{E} \times \overline{\boldsymbol{H}}]$ is real and constant with respect to the time variable t. The second term is also real but varying in time with frequency 2ω, so that its time average is zero. Hence the time average of $\boldsymbol{\mathcal{E}} \times \boldsymbol{\mathcal{H}}$ is equal to the real part of $\boldsymbol{S}$ where

$$\boldsymbol{S} := \frac{1}{2}[\boldsymbol{E} \times \overline{\boldsymbol{H}}] \tag{2.17}$$

denotes the **complex Poynting vector.**

The time average of the power flux from the volume into the region outside is then given by (cf. (2.12))

$$P = \operatorname{Re} \int_{\partial\Omega} \boldsymbol{S} \cdot \boldsymbol{n}\, dS. \tag{2.18}$$

2.8 Vector Potentials

The advantage of the vector Helmholtz equations (2.16) over the original Maxwell system (2.13a)–(2.13d) is that every Cartesian component u of the fields satisfies the scalar Helmholtz equation $\Delta u + k^2 u = f$ where f denotes the corresponding component of the right hand side. However, since the condition on the divergence couples the components again, it is not an easy task to construct solutions of (2.13a)- (2.13d) or, equivalently, (2.16) directly. This is the main reason why it is very convenient to introduce **vector potentials $\boldsymbol{A}$.** In this section, we assume that the medium is homogeneous, i.e. ϵ, μ and σ are constant.

It is well known (see, e.g., [94]) that in regions Ω without interior boundaries the condition $\operatorname{div}\boldsymbol{H} = 0$ is equivalent to the existence of a vector field $\boldsymbol{A}$ such that $\boldsymbol{H} = \operatorname{curl}\boldsymbol{A}$. Vector fields $\boldsymbol{A}$ with this property are called *vector potentials* for the field $\boldsymbol{H}$. Note that they are not unique. Indeed, with $\boldsymbol{A}$ also $\boldsymbol{A} + \nabla\varphi$ for any differentiable function φ is also a vector potential for $\boldsymbol{H}$. Substitution of $\boldsymbol{H} = \operatorname{curl}\boldsymbol{A}$ into the second equation of (2.16) yields

$$\operatorname{curl}\left(\Delta\boldsymbol{A} + k^2\boldsymbol{A}\right) = -\operatorname{curl}\boldsymbol{J}_e$$

which is certainly satisfied if

$$\Delta\boldsymbol{A} + k^2\boldsymbol{A} = -\boldsymbol{J}_e + \nabla\varphi \tag{2.19}$$

where φ is any differentiable scalar function.

On the other hand, if $\boldsymbol{A}$ satisfies (2.19) then

$$\boldsymbol{H} \;=\; \operatorname{curl}\boldsymbol{A} \quad \text{and} \quad \boldsymbol{E} \;=\; i\omega\mu_0\,\boldsymbol{A} \;+\; \frac{1}{\sigma - i\omega\epsilon}\nabla(\operatorname{div}\boldsymbol{A} - \varphi) \tag{2.20}$$

satisfies the Maxwell system (2.13a)–(2.13d) with $\rho = \operatorname{div}\boldsymbol{J}_e/(i\omega - \sigma/\epsilon)$. The vector potential $\boldsymbol{A}$ used to express the magnetic field $\boldsymbol{H}$ is also called **magnetic Hertz potential**. The following example is of particular importance:

Example 2.1. (TM-mode)

Let $\boldsymbol{A}$ be a solution of (2.19) of the form $\boldsymbol{A}(\boldsymbol{x}) = u(\boldsymbol{x})\,\boldsymbol{p}(\boldsymbol{x})$ with scalar field u and vector field $\boldsymbol{p}$ such that $\operatorname{curl}\boldsymbol{p} = \boldsymbol{o}$. This situation is called *TM-mode* (transverse-magnetic mode) since $\boldsymbol{H} = \operatorname{curl}(u\,\boldsymbol{p}) = \nabla u \times \boldsymbol{p}$ has no component in $\boldsymbol{p}-$direction.

As a particular example we take $\boldsymbol{p}$ being constant, without loss of generality $\boldsymbol{p} = \hat{\boldsymbol{e}}_3$, the unit vector in x_3-direction, and $\boldsymbol{J}_e = g\,\hat{\boldsymbol{e}}_3$. If u is a solution of the **three dimensional scalar Helmholtz equation**

$$\Delta u \;+\; k^2 u \;=\; -g \tag{2.21}$$

we have

$$\boldsymbol{H} = \operatorname{curl}(u\,\hat{\boldsymbol{e}}_3) \;=\; \left(\frac{\partial u}{\partial x_2}\,,\, -\frac{\partial u}{\partial x_1}\,,\, 0\right)^{\top}, \tag{2.22a}$$

$$\boldsymbol{E} = i\omega\mu_0\,u\,\hat{\boldsymbol{e}}_3 \;+\; \frac{1}{\sigma - i\omega\epsilon}\,\nabla\big(\partial u/\partial x_3\big)\,. \tag{2.22b}$$

If we choose g and u to be constant with respect to x_3 then $\boldsymbol{E}$ has only a x_3-component. This mode is also called **E-mode**. In this case equation (2.21) reduces to the two dimensional scalar Helmholtz equation for u.

Analogously, we can introduce **electric Hertz potentials**. Indeed, if $\boldsymbol{J}_e = \boldsymbol{o}$ then $\operatorname{div}\boldsymbol{E} = 0$ and we substitute the ansatz $\boldsymbol{E} = \operatorname{curl}\boldsymbol{A}$ into the first equation of (2.16). This yields

$$\operatorname{curl}\big(\Delta\boldsymbol{A} \;+\; k^2\boldsymbol{A}\big) \;=\; \boldsymbol{o}$$

which is certainly satisfied if

$$\Delta\boldsymbol{A} \;+\; k^2\boldsymbol{A} \;=\; \nabla\varphi \tag{2.23}$$

where again φ is any differentiable scalar function. If, on the other hand, $\boldsymbol{A}$ satisfies equation (2.23) then

$$\boldsymbol{E} \;=\; \operatorname{curl}\boldsymbol{A} \quad \text{and} \quad \boldsymbol{H} \;=\; (\sigma - i\omega\epsilon)\,\boldsymbol{A} \;+\; \frac{1}{i\omega\mu_0}\nabla(\operatorname{div}\boldsymbol{A} - \varphi) \tag{2.24}$$

satisfies the Maxwell system (2.13a)–(2.13d). Analogously to above we consider the following example:

Example 2.2. (TE-mode)
Let $\boldsymbol{J}_e = \boldsymbol{o}$ and $\boldsymbol{A}$ be a solution of (2.23) of the form $\boldsymbol{A} = u\,\boldsymbol{p}$ with $\operatorname{curl}\boldsymbol{p} = \boldsymbol{o}$. This situation describes the **TE-mode** since now $\boldsymbol{E}$ has no component in $\boldsymbol{p}$–direction. With the particular choice $\boldsymbol{p} = \hat{\boldsymbol{e}}_3$ and a scalar solution u of the three dimensional Helmholtz equation $\Delta u + k^2 u = 0$ we have

$$\boldsymbol{E} = \operatorname{curl}(\hat{\boldsymbol{e}}_3 u) \;=\; \left(\frac{\partial u}{\partial x_2}, -\frac{\partial u}{\partial x_1}, 0 \right)^\top , \qquad (2.25a)$$

$$\boldsymbol{H} = (\sigma - i\omega\epsilon)\, u\, \hat{\boldsymbol{e}}_3 \;+\; \frac{1}{i\omega\mu_0}\, \nabla\big(\partial u/\partial x_3\big)\,. \qquad (2.25b)$$

The case where u is independent of x_3 is also called **H-mode**.

We observe that in both, the H- and the E-mode the electric and magnetic fields are perpendicular to each other. This is not true, in general, for the TE or TM mode or even for arbitrary solutions of Maxwell's equations (except in the far field, cf. (2.32)).

2.9 Radiation Condition, Far Field Pattern

We will see that solutions of Maxwell's equations decay or increase exponentially for conducting media, i.e. when $\sigma > 0$. For $\sigma = 0$, however, every solution must decay as $1/r$ for $r \to \infty$. To illustrate this let us consider one of the simplest possible magnetic Hertz potentials, namely those which are radially symmetric. Therefore, let us assume that $\mu = \mu_0$, ϵ, and σ are constant, $\boldsymbol{A} = \boldsymbol{A}(r)$, $r > 0$, and $\boldsymbol{J}_e = \boldsymbol{o}$, $\varphi = 0$. The Helmholtz equation (2.19) in spherical coordinates reduces to the ordinary differential equation

$$\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2 \frac{\partial \boldsymbol{A}(r)}{\partial r} \right) \;+\; k^2\, \boldsymbol{A}(r) \;=\; \boldsymbol{o}\,, \quad r > 0\,, \qquad (2.26)$$

which has the two linearly independent solutions

$$\boldsymbol{A}(r) \;=\; \boldsymbol{p}\,\frac{\mathrm{e}^{ikr}}{r} \quad \text{and} \quad \boldsymbol{A}(r) \;=\; \boldsymbol{p}\,\frac{\mathrm{e}^{-ikr}}{r}$$

as it is readily seen. The corresponding magnetic- and electric fields are given by (2.20), i.e.

$$\boldsymbol{H}(\boldsymbol{x}) = \operatorname{curl}\boldsymbol{A}(\boldsymbol{x}) \;=\; \nabla\frac{\mathrm{e}^{\pm ikr}}{r} \times \boldsymbol{p}\,,$$

$$\boldsymbol{E}(\boldsymbol{x}) = i\omega\mu_0\,\boldsymbol{p}\,\frac{\mathrm{e}^{\pm ikr}}{r} \;+\; \frac{1}{\sigma - i\omega\epsilon}\,\nabla\left(\boldsymbol{p}\cdot\nabla\frac{\mathrm{e}^{\pm ikr}}{r}\right).$$

From the asymptotic behaviour

$$\nabla \frac{e^{\pm ikr}}{r} = \pm ik \, \frac{e^{\pm ikr}}{r} \, [\hat{\boldsymbol{x}} + \mathcal{O}(1/r)] \quad \text{as } r \to \infty ,$$

$$\frac{\partial^2}{\partial x_j \partial x_\ell} \frac{e^{\pm ikr}}{r} = -k^2 \, \frac{e^{\pm ikr}}{r} \left[\frac{x_j}{r} \frac{x_\ell}{r} + \mathcal{O}(1/r) \right] \quad \text{as } r \to \infty ,$$

uniformly in $\hat{\boldsymbol{x}} \in S^2$, we observe that

$$\boldsymbol{H}(\boldsymbol{x}) = \pm ik \, \frac{e^{\pm ikr}}{r} \, [\hat{\boldsymbol{x}} \times \boldsymbol{p} + \mathcal{O}(1/r)] , \tag{2.27a}$$

$$\begin{aligned} \boldsymbol{E}(\boldsymbol{x}) &= i\omega\mu_0 \, \frac{e^{\pm ikr}}{r} \, [\boldsymbol{p} - \hat{\boldsymbol{x}}(\hat{\boldsymbol{x}} \cdot \boldsymbol{p}) + \mathcal{O}(1/r)] \\ &= i\omega\mu_0 \, \frac{e^{\pm ikr}}{r} \, [\hat{\boldsymbol{x}} \times (\boldsymbol{p} \times \hat{\boldsymbol{x}}) + \mathcal{O}(1/r)] , \end{aligned} \tag{2.27b}$$

$$\boldsymbol{S}(\boldsymbol{x}) = \frac{1}{2} \, \boldsymbol{E}(\boldsymbol{x}) \times \overline{\boldsymbol{H}(\boldsymbol{x})} \tag{2.27c}$$

$$= \pm \frac{1}{2} \, \omega\mu_0 \, \overline{k} \, \frac{e^{\mp 2r \operatorname{Im} k}}{r^2} \left[\left(|\boldsymbol{p}|^2 - |\hat{\boldsymbol{x}} \cdot \boldsymbol{p}|^2 \right) \hat{\boldsymbol{x}} + \mathcal{O}(1/r) \right] . \tag{2.27d}$$

Here we clearly see the asymptotic behavior as $r \to \infty$: for $\sigma > 0$ we have that $\operatorname{Im} k > 0$, i.e. $\boldsymbol{H}$, $\boldsymbol{E}$, and $\boldsymbol{S}$ are exponentially decreasing or increasing, respectively, depending on the sign in the exponential term. If $\sigma = 0$, however, k is real valued and the fields $\boldsymbol{E}$ and $\boldsymbol{H}$ decay as $1/r$ while $\boldsymbol{S}$ decays as $1/r^2$. This is different from the static case where it is well known that the fields could decay more rapidly (see also Section 5.2).

We now formulate radiation conditions on $\boldsymbol{E}$ and $\boldsymbol{H}$ which are independent of the special example for $\boldsymbol{A}$ and distinguish between the two possible solutions.

If the medium is conducting i.e. if $\operatorname{Im} k > 0$, then, from conservation of energy, the radiated power cannot increase with r, thus we must take the positive sign in the exponential terms of $\boldsymbol{A}$, $\boldsymbol{E}$ and $\boldsymbol{H}$. Formulated in terms of $\boldsymbol{E}$ and $\boldsymbol{H}$ it is sufficient to require that

$$\boldsymbol{E} \quad \text{and} \quad \boldsymbol{H} \quad \text{are bounded.} \tag{2.28}$$

In vacuo, $\epsilon = \epsilon_0$ and $\sigma = 0$, i.e. k is real valued and positive. We observe that, by using the Cauchy-Schwarz inequality, the Poynting vector

$$\boldsymbol{S}(\boldsymbol{x}) = \frac{1}{2} \, \omega\mu_0 \, k \, \frac{1}{r^2} \left[\left(|\boldsymbol{p}|^2 - |\hat{\boldsymbol{x}} \cdot \boldsymbol{p}|^2 \right) \hat{\boldsymbol{x}} + \mathcal{O}(1/r) \right]$$

is directed into the direction $\hat{\boldsymbol{x}}$ which represents outgoing rather than incoming fields. Therefore, we also choose the positive sign in the exponential terms of $\boldsymbol{A}$, $\boldsymbol{E}$ and $\boldsymbol{H}$. Then, $\boldsymbol{E}$ and $\boldsymbol{H}$ satisfy the **Silver-Müller radiation conditions**:

$$\boldsymbol{E}(\boldsymbol{x}) \times \hat{\boldsymbol{x}} + \frac{1}{Y_0} \, \boldsymbol{H}(\boldsymbol{x}) = \mathcal{O}\left(1/r^2\right) , \tag{2.29a}$$

$$\boldsymbol{H}(\boldsymbol{x}) \times \hat{\boldsymbol{x}} - Y_0 \, \boldsymbol{E}(\boldsymbol{x}) = \mathcal{O}\left(1/r^2\right) \tag{2.29b}$$

as $r \to \infty$ uniformly with respect to $\hat{\boldsymbol{x}} \in S^2$. Here,

$$Y_0 := \sqrt{\frac{\epsilon_0}{\mu_0}} = \frac{k}{\omega\mu_0} = \frac{\omega\epsilon_0}{k} \tag{2.30}$$

denotes the *admittance* in non-conductive media. In vacuo it is $Y_0 \approx 2.654 \cdot 10^{-3}\, A/V$.

It turns out that these radiation conditions describe the correct asymptotic behavior of electromagnetic waves generated by sources lying in a *compact set.* It can be shown (see [29]) that these conditions are equivalent (i.e. any solution $(\boldsymbol{E}, \boldsymbol{H})$ of the time harmonic Maxwell's equations which satisfies one of (2.29a), (2.29b) also satisfies the other one) and that they imply the following asymptotic behavior of $\boldsymbol{E}$ and $\boldsymbol{H}$:

$$\boldsymbol{E}(\boldsymbol{x}) = \frac{\mathrm{e}^{ikr}}{r}\left[\boldsymbol{E}_\infty(\hat{\boldsymbol{x}}) \;+\; \mathcal{O}(1/r)\right], \tag{2.31a}$$

$$\boldsymbol{H}(x) = Y_0\,\frac{\mathrm{e}^{ikr}}{r}\left[\boldsymbol{H}_\infty(\hat{\boldsymbol{x}}) \;+\; \mathcal{O}(1/r)\right] \tag{2.31b}$$

as $r \to \infty$ uniformly with respect to $\hat{\boldsymbol{x}} \in S^2$. The vector fields $\boldsymbol{E}_\infty$ and $\boldsymbol{H}_\infty$ are defined on the unit sphere S^2 and are called *far field pattern.* In the particular example above the far field patterns are given by

$$\boldsymbol{H}_\infty(\hat{\boldsymbol{x}}) \;=\; \frac{ik}{Y_0}\,\hat{\boldsymbol{x}} \times \boldsymbol{p} \;=\; i\omega\mu_0\,\hat{\boldsymbol{x}} \times \boldsymbol{p}\,, \qquad \boldsymbol{E}_\infty(\hat{\boldsymbol{x}}) \;=\; i\omega\mu_0\,\hat{\boldsymbol{x}} \times (\boldsymbol{p} \times \hat{\boldsymbol{x}})\,.$$

In general, they enjoy the following properties (cf. [29]):

$$\boldsymbol{H}_\infty(\hat{\boldsymbol{x}}) \;=\; \hat{\boldsymbol{x}}\times\boldsymbol{E}_\infty(\hat{\boldsymbol{x}}), \quad \hat{\boldsymbol{x}}\cdot\boldsymbol{E}_\infty(\hat{\boldsymbol{x}}) \;=\; \hat{\boldsymbol{x}}\cdot\boldsymbol{H}_\infty(\hat{\boldsymbol{x}}) \;=\; 0 \quad \text{for } \hat{\boldsymbol{x}} \in S^2. \tag{2.32}$$

To explain the physical meaning of the far field pattern we consider the energy distribution which is given by the complex Poynting vector $\frac{1}{2}[\boldsymbol{E} \times \overline{\boldsymbol{H}}]$ (cf. (2.17)) in any nonconducting medium, i.e. we assume k and Y_0 are both real valued and positive. The time averaged power radiated through the sphere S_a of radius a can be written as (cf. (2.12)):

$$P_a \;=\; \mathrm{Re}\left[\int_{S_a} \boldsymbol{n}\cdot\boldsymbol{S}\,dS\right] \;=\; \frac{1}{2}\,\mathrm{Re}\left[\int_{S_a} \hat{\boldsymbol{x}}\cdot(\boldsymbol{E}\times\overline{\boldsymbol{H}})\,dS\right], \tag{2.33}$$

and so the power radiated into the far field is given by

$$P_\infty \;:=\; \frac{1}{2}\,\mathrm{Re}\left[\lim_{a\to\infty}\int_{S_a} \hat{\boldsymbol{x}}\cdot(\boldsymbol{E}\times\overline{\boldsymbol{H}})\,dS\right]. \tag{2.34}$$

Using the definitions of the far field patterns (2.31a), (2.31b) and the properties (2.32) we can express P_∞ in terms of $\boldsymbol{E}_\infty$ alone:

$$P_\infty \;=\; \frac{Y_0}{2}\int_{S^2} |\boldsymbol{E}_\infty|^2\,dS\,. \tag{2.35}$$

2.10 Radiating Dipoles and Line Sources

The construction of solutions of the Maxwell equations (2.13a)–(2.13d) by introducing vector potentials is a purely mathematical approach. In this section we will briefly connect this construction with the physical electromagnetic fields radiated by infinitesimal or finite linear current elements.

To derive the fields for an electric dipole we start with a small volume element $\Delta V(\boldsymbol{z})$ centered at $\boldsymbol{z}$ which e.g., can be, but doesn't have to be, a ball with center $\boldsymbol{z}$ and radius ϵ. Assuming the total current to be $I\,\hat{\boldsymbol{a}}$ for some unit vector $\hat{\boldsymbol{a}}$ we define the current density $\boldsymbol{J}_e$ by $\boldsymbol{J}_e(\boldsymbol{x}) = j(\boldsymbol{x})\,\hat{\boldsymbol{a}}$ where $j \in C^1(\mathbb{R}^3)$ is any function with the properties that $j(\boldsymbol{x}) = 0$ for $\boldsymbol{x} \notin \Delta V(\boldsymbol{z})$ and $\int_{\Delta V(\boldsymbol{z})} j(\boldsymbol{y})\, d\boldsymbol{y} = I$.

We solve Maxwell's equations (2.13a)–(2.13d) for this particular current distribution by introducing a magnetic Hertz potential $\boldsymbol{A} = u\,\hat{\boldsymbol{a}}$. Then u must satisfy the Helmholtz equation

$$\Delta u + k^2 u = -j \quad \text{in } \mathbb{R}^3 . \tag{2.36}$$

The following theorem is well known (see, e.g., [41])

Theorem 2.3. *Let $\Omega \subset \mathbb{R}^3$ be a bounded domain and*

$$\Phi(\boldsymbol{x},\boldsymbol{y}) := \frac{e^{ik|\boldsymbol{x}-\boldsymbol{y}|}}{4\pi\,|\boldsymbol{x}-\boldsymbol{y}|}, \quad \boldsymbol{x} \neq \boldsymbol{y}, \tag{2.37}$$

denote the fundamental solution *of the Helmholtz equation in $\mathbb{R}^3$. Then, for every $j \in C^1(\overline{\Omega})$, the volume potential*

$$u(\boldsymbol{x}) = \int_{\Omega} \Phi(\boldsymbol{x},\boldsymbol{y})\, j(\boldsymbol{y})\, d\boldsymbol{y}, \quad \boldsymbol{x} \in \mathbb{R}^3, \tag{2.38}$$

with density j is two times continuously differentiable in Ω and in $\mathbb{R}^3 \setminus \overline{\Omega}$ and

$$\Delta u + k^2 u = -j \text{ in } \Omega, \quad \Delta u + k^2 u = 0 \text{ in } \mathbb{R}^3 \setminus \overline{\Omega}.$$

Furthermore, u satisfies the **Sommerfeld radiation condition**

$$\frac{\partial}{\partial r} u(\boldsymbol{x}) - i\,k\,u(\boldsymbol{x}) = \mathcal{O}\big(1/r^2\big) \quad \text{as } r \to \infty, \tag{2.39}$$

uniformly with respect to $\hat{\boldsymbol{x}} = \boldsymbol{x}/|\boldsymbol{x}| \in S^2$.

Applying this result to the current distribution yields that u, defined by

$$\begin{aligned} u(\boldsymbol{x}) &= \int_{\Delta V(\boldsymbol{z})} \Phi(\boldsymbol{x},\boldsymbol{y})\, j(\boldsymbol{y})\, d\boldsymbol{y} \\ &= I\,\Phi(\boldsymbol{x},\boldsymbol{z}) + \int_{\Delta V(\boldsymbol{z})} j(\boldsymbol{y})\,\big[\Phi(\boldsymbol{x},\boldsymbol{y}) - \Phi(\boldsymbol{x},\boldsymbol{z})\big]\, d\boldsymbol{y}, \end{aligned}$$

solves (2.36). For fixed $\boldsymbol{x} \neq \boldsymbol{z}$ we let the region $\Delta V(\boldsymbol{z})$ shrink to the point $\boldsymbol{z}$ while keeping the total current I fixed. Then $u(\boldsymbol{x})$ converges to $u(\boldsymbol{x}) = I\,\Phi(\boldsymbol{x},\boldsymbol{z})$.

Remarks:

- Actually, the function $\Phi_-(\boldsymbol{x},\boldsymbol{y}) := \frac{\exp(-ik|\boldsymbol{x}-\boldsymbol{y}|)}{4\pi|\boldsymbol{x}-\boldsymbol{y}|}$ is also a fundamental solution. As we have made clear in the previous section, however, the fields based on potentials with Φ_- do not satisfy the radiation condition and are therefore physically not relevant.
- By this "shrinking process" the current density j has to tend to infinity since the total current I is kept fixed. The limit of these currents j cannot be a function in the ordinary sense. Therefore, by this limiting process, we extend the concept of a function to the wider class of " distributions". We actually take $j(\boldsymbol{x}) = I\,\delta(\boldsymbol{x}-\boldsymbol{z})$ where δ denotes Dirac's delta-distribution introduced formally by the property $\int_{\mathbb{R}^3} \delta(\boldsymbol{y})g(\boldsymbol{y})\,d\boldsymbol{y} = g(0)$ for every $g \in C(\mathbb{R}^3)$. We can write formally

$$\Delta_x \Phi(\boldsymbol{x},\boldsymbol{z}) \;+\; k^2\Phi(\boldsymbol{x},\boldsymbol{z}) \;=\; -\delta(\boldsymbol{x}-\boldsymbol{z}).$$

  This formulation can be made mathematically rigorous by using the theory of distributions (see [136]).

The magnetic and electric fields corresponding to the potential $\boldsymbol{A}(\boldsymbol{x}) = I\,\Phi(\boldsymbol{x},\boldsymbol{z})\,\hat{\boldsymbol{a}}$ are given by (2.20), i.e.

$$\boldsymbol{H}(\boldsymbol{x}) = I\,\mathrm{curl}\,\big(\Phi(\boldsymbol{x},\boldsymbol{z})\,\hat{\boldsymbol{a}}\big) \;=\; I\,\nabla_x\Phi(\boldsymbol{x},\boldsymbol{z}) \times \hat{\boldsymbol{a}} \tag{2.40a}$$

$$\boldsymbol{E}(\boldsymbol{x}) = i\omega\mu_0\,I\,\Phi(\boldsymbol{x},\boldsymbol{z})\,\hat{\boldsymbol{a}} \;+\; \frac{I}{\sigma - i\omega\epsilon}\,\nabla_x\Big(\nabla_x\Phi(\boldsymbol{x},\boldsymbol{z})\cdot\hat{\boldsymbol{a}}\Big)\,. \tag{2.40b}$$

These are the fields of an **electric dipole** with dipole moment $I\hat{\boldsymbol{a}}$.

We now want to derive the fields of a **magnetic dipole**. Let again $\Delta V(\boldsymbol{z})$ be a volume element with center $\boldsymbol{z}$ and $j_m(\boldsymbol{x})$ as before an approximation of $M\,\delta(\boldsymbol{x})$. The vector field $\boldsymbol{J}_e(\boldsymbol{x}) = \mathrm{curl}\,(j_m(\boldsymbol{x})\,\hat{\boldsymbol{a}}) = \nabla j_m(\boldsymbol{x}) \times \hat{\boldsymbol{a}}$ describes the current distribution of a small *circular wire.* We set $\boldsymbol{J}_m := j_m(\boldsymbol{x})\,\hat{\boldsymbol{a}}$ and call this auxiliary quantity a **magnetic current distribution**. It is our aim to solve Maxwell's equations (2.13a)- (2.13d) with this choice of $\boldsymbol{J}_e = \mathrm{curl}\,\boldsymbol{J}_m$. Instead of going through the details again we make use of a mathematical trick:

We define the purely mathematical vector fields

$$\tilde{\boldsymbol{E}} \;:=\; \boldsymbol{H} \;-\; \boldsymbol{J}_m \quad \text{and} \quad \tilde{\boldsymbol{H}} \;:=\; \boldsymbol{E}\,.$$

Then, since $\mathrm{div}\,\boldsymbol{J}_e = 0$, the fields $\tilde{\boldsymbol{E}}$, $\tilde{\boldsymbol{H}}$ solve the system

$$\mathrm{curl}\,\tilde{\boldsymbol{E}} = (\sigma - i\omega\epsilon)\,\tilde{\boldsymbol{H}}\,, \quad \mathrm{curl}\,\tilde{\boldsymbol{H}} = i\omega\mu_0\,\tilde{\boldsymbol{E}} + i\omega\mu_0\boldsymbol{J}_m\,,$$

$$\operatorname{div}\tilde{\boldsymbol{E}} = -\operatorname{div}\boldsymbol{J}_m\,, \quad \operatorname{div}\tilde{\boldsymbol{H}} = 0\,.$$

Formally, this looks like a Maxwell system with the roles of $\sigma - i\omega\epsilon$ and $i\omega\mu_0$ interchanged. The current density in this case is $i\omega\mu_0\boldsymbol{J}_m$. Therefore, we solve this system as in the case of an electric dipole and arrive at the potential $\boldsymbol{A}(\boldsymbol{x}) = i\omega\mu_0 M\,\Phi(\boldsymbol{x},\boldsymbol{z})\,\hat{\boldsymbol{a}}$ and thus:

$$\boldsymbol{E}(\boldsymbol{x}) = \operatorname{curl}\boldsymbol{A}(\boldsymbol{x}) \;=\; i\omega\mu_0\,M\,\nabla_x\Phi(\boldsymbol{x},\boldsymbol{z})\times\hat{\boldsymbol{a}}\,, \tag{2.41a}$$

$$\begin{aligned}\boldsymbol{H}(\boldsymbol{x}) &= \frac{1}{i\omega\,\mu}\operatorname{curl}\boldsymbol{E}(\boldsymbol{x}) \;=\; M\left(-\Delta + \nabla\operatorname{div}\right)\left(\Phi(\boldsymbol{x},\boldsymbol{z})\hat{\boldsymbol{a}}\right)\\ &= k^2 M\,\Phi(\boldsymbol{x},\boldsymbol{z})\,\hat{\boldsymbol{a}} \;+\; M\,\nabla_x\Big(\nabla_x\Phi(\boldsymbol{x},\boldsymbol{z})\cdot\hat{\boldsymbol{a}}\Big)\,.\end{aligned} \tag{2.41b}$$

These are the fields of a **magnetic dipole** with dipole moment $M\hat{\boldsymbol{a}}$.

To find the asymptotic behaviour of these fields and the corresponding far field patterns we have to study the fundamental solution $\Phi(\boldsymbol{x},\boldsymbol{z})$ as $r = |\boldsymbol{x}|$ tends to infinity. From the representation

$$|\boldsymbol{x}-\boldsymbol{z}| \;=\; |\boldsymbol{x}| \;-\; \frac{\boldsymbol{x}\cdot\boldsymbol{z}}{|\boldsymbol{x}|} \;+\; a(\boldsymbol{x},\boldsymbol{z}) \quad\text{with}\quad |a(\boldsymbol{x},\boldsymbol{z})| \;\le\; 4\,\frac{|\boldsymbol{z}|^2}{|\boldsymbol{x}|}$$

for all $\boldsymbol{x},\boldsymbol{z}\in\mathbb{R}^3$ with $\boldsymbol{x}\neq\boldsymbol{o}$, $|\boldsymbol{z}|\le\frac{1}{2}|\boldsymbol{x}|$, we derive the asymptotic representation of the fundamental solution Φ of the Helmholtz equation in the form (using polar coordinates r, θ and ϕ with respect to the origin):

$$\Phi(\boldsymbol{x},\boldsymbol{z}) = \frac{\mathrm{e}^{ikr}}{4\pi r}\left[\mathrm{e}^{-ik\hat{\boldsymbol{x}}\cdot\boldsymbol{z}} \;+\; \mathcal{O}(1/r)\right] \quad \text{as } r = |\boldsymbol{x}|\to\infty, \tag{2.42a}$$

$$\nabla_x\Phi(\boldsymbol{x},\boldsymbol{z}) = ik\,\frac{\mathrm{e}^{ikr}}{4\pi r}\left[\mathrm{e}^{-ik\hat{\boldsymbol{x}}\cdot\boldsymbol{z}}\,\hat{\boldsymbol{x}} \;+\; \mathcal{O}(1/r)\right] \quad \text{as } r\to\infty, \tag{2.42b}$$

$$\frac{\partial^2\Phi(\boldsymbol{x},\boldsymbol{z})}{\partial x_j\partial x_\ell} = -k^2\,\frac{\mathrm{e}^{ikr}}{4\pi r}\left[\mathrm{e}^{-ik\hat{\boldsymbol{x}}\cdot\boldsymbol{z}}\frac{x_j}{r}\,\frac{x_\ell}{r} \;+\; \mathcal{O}(1/r)\right] \quad \text{as } r\to\infty\,, \tag{2.42c}$$

uniformly in θ, ϕ and $\boldsymbol{z}$ in any compact subset of $\mathbb{R}^3$. Again, we have set $\hat{\boldsymbol{x}} = \boldsymbol{x}/r$. From this we see that the fields generated by an **electric dipole** or **magnetic dipole** with moment $I\hat{\boldsymbol{a}}$ and $M\hat{\boldsymbol{a}}$, respectively, satisfy

$$\boldsymbol{H}(\boldsymbol{x}) = ik\,I\,\frac{\mathrm{e}^{ikr}}{4\pi r}\left[\mathrm{e}^{-ik\hat{\boldsymbol{x}}\cdot\boldsymbol{z}}(\hat{\boldsymbol{x}}\times\hat{\boldsymbol{a}}) \;+\; \mathcal{O}(1/r)\right], \tag{2.43a}$$

$$\boldsymbol{E}(\boldsymbol{x}) = i\omega\mu_0\,I\,\frac{\mathrm{e}^{ikr}}{4\pi r}\left[\mathrm{e}^{-ik\hat{\boldsymbol{x}}\cdot\boldsymbol{z}}[\hat{\boldsymbol{x}}\times(\hat{\boldsymbol{a}}\times\hat{\boldsymbol{x}})] \;+\; \mathcal{O}(1/r)\right] \tag{2.43b}$$

and

$$\boldsymbol{E}(\boldsymbol{x}) = -k\omega\mu_0\,M\,\frac{\mathrm{e}^{ikr}}{4\pi r}\left[\mathrm{e}^{-ik\hat{\boldsymbol{x}}\cdot\boldsymbol{z}}(\hat{\boldsymbol{x}}\times\hat{\boldsymbol{a}}) \;+\; \mathcal{O}(1/r)\right], \tag{2.43c}$$

$$\boldsymbol{H}(\boldsymbol{x}) = k^2\,M\,\frac{\mathrm{e}^{ikr}}{4\pi r}\left[\mathrm{e}^{-ik\hat{\boldsymbol{x}}\cdot\boldsymbol{z}}[\hat{\boldsymbol{x}}\times(\hat{\boldsymbol{a}}\times\hat{\boldsymbol{x}})] \;+\; \mathcal{O}(1/r)\right], \tag{2.43d}$$

respectively, as $r \to \infty$.

The far field patterns of an electric dipole have been computed in Section 2.9 already. (This was the motivation for the Silver-Müller radiation condition). We repeat that the far field patterns generated by an electric or magnetic dipole are given by

$$\boldsymbol{H}_\infty(\hat{\boldsymbol{x}}) = \frac{ikI}{4\pi}\, \mathrm{e}^{-ik\hat{\boldsymbol{x}}\cdot\boldsymbol{z}}\, (\hat{\boldsymbol{x}} \times \hat{\boldsymbol{a}})\,, \tag{2.44a}$$

$$\boldsymbol{E}_\infty(\hat{\boldsymbol{x}}) = \frac{i\omega\mu_0 I}{4\pi}\, \mathrm{e}^{-ik\hat{\boldsymbol{x}}\cdot\boldsymbol{z}}\, [\hat{\boldsymbol{x}} \times (\hat{\boldsymbol{a}} \times \hat{\boldsymbol{x}})]\,. \tag{2.44b}$$

and

$$\boldsymbol{E}_\infty(\hat{\boldsymbol{x}}) = \frac{-k\omega\mu_0 M}{4\pi}\, \mathrm{e}^{-ik\hat{\boldsymbol{x}}\cdot\boldsymbol{z}}\, (\hat{\boldsymbol{a}} \times \hat{\boldsymbol{x}})\,, \tag{2.44c}$$

$$\boldsymbol{H}_\infty(\hat{\boldsymbol{x}}) = \frac{k^2 M}{4\pi}\, \mathrm{e}^{-ik\hat{\boldsymbol{x}}\cdot\boldsymbol{z}}\, [\hat{\boldsymbol{x}} \times (\hat{\boldsymbol{a}} \times \hat{\boldsymbol{x}})]\,. \tag{2.44d}$$

Example 2.4. As a special example we consider the case $\hat{\boldsymbol{a}} = \hat{\boldsymbol{e}}_3$ (the unit vector in x_3–direction). Using spherical polar coordinates (r, θ, ϕ) of $\boldsymbol{x}$ with respect to $\boldsymbol{z}$ and coordinate vectors $\hat{\boldsymbol{x}}$, $\hat{\boldsymbol{\theta}}$, $\hat{\boldsymbol{\phi}}$ we compute

$$\nabla_x \Phi(\boldsymbol{x}, \boldsymbol{z}) = \left(ik - \frac{1}{r}\right) \frac{\mathrm{e}^{ikr}}{4\pi r}\, \hat{\boldsymbol{x}}, \tag{2.45a}$$

$$\nabla_x \Phi(\boldsymbol{x}, \boldsymbol{z}) \times \hat{\boldsymbol{e}}_3 = -\left(ik - \frac{1}{r}\right) \frac{\mathrm{e}^{ikr}}{4\pi r}\, \sin\theta\, \hat{\boldsymbol{\phi}} \tag{2.45b}$$

(since $\hat{\boldsymbol{x}} \times \hat{\boldsymbol{e}}_3 = -\sin\theta\, \hat{\boldsymbol{\phi}}$), and

$$\nabla_x \frac{\partial}{\partial x_3} \Phi(\boldsymbol{x}, \boldsymbol{z}) = \left(\frac{3}{r^2} - \frac{3ik}{r} - k^2\right) \frac{\mathrm{e}^{ikr}}{4\pi r}\, \cos\theta\, \hat{\boldsymbol{x}} + \left(\frac{ik}{r} - \frac{1}{r^2}\right) \frac{\mathrm{e}^{ikr}}{4\pi r}\, \hat{\boldsymbol{e}}_3 \tag{2.45c}$$

and thus for the *electric dipole* by (2.40a), (2.40b):

$$\boldsymbol{H}(\boldsymbol{x}) = -I \left(ik - \frac{1}{r}\right) \frac{\mathrm{e}^{ikr}}{4\pi r}\, \sin\theta\, \hat{\boldsymbol{\phi}}, \tag{2.46a}$$

$$\boldsymbol{E}(\boldsymbol{x}) = i\omega\mu_0 I\, \Phi(\boldsymbol{x}, \boldsymbol{z}) \hat{\boldsymbol{e}}_3 \;-\; \frac{I}{\sigma - i\omega\epsilon}\, \nabla \frac{\partial}{\partial x_3} \Phi(\boldsymbol{x}, \boldsymbol{z})$$

i.e., since $\hat{\boldsymbol{e}}_3 = \cos\theta\, \hat{\boldsymbol{x}} - \sin\theta\, \hat{\boldsymbol{\theta}}$,

$$\boldsymbol{E}(\boldsymbol{x}) = \frac{2I}{\sigma - i\omega\epsilon} \left(\frac{1}{r} - ik\right) \frac{\mathrm{e}^{ikr}}{4\pi r^2}\, \cos\theta\, \hat{\boldsymbol{x}}$$

$$- \frac{I}{\sigma - i\omega\epsilon} \left(k^2 + \frac{ik}{r} - \frac{1}{r^2}\right) \frac{\mathrm{e}^{ikr}}{4\pi r}\, \sin\theta\, \hat{\boldsymbol{\theta}}. \tag{2.46b}$$

Analogously, we have for the *magnetic dipole*

$$\boldsymbol{E}(\boldsymbol{x}) = -i\omega\,\mu\,M\left(ik - \frac{1}{r}\right)\frac{\mathrm{e}^{ikr}}{4\pi r}\,\sin\theta\,\hat{\boldsymbol{\phi}}, \tag{2.47a}$$

$$\boldsymbol{H}(\boldsymbol{x}) = 2M\left(\frac{1}{r} - ik\right)\frac{\mathrm{e}^{ikr}}{4\pi r^2}\,\cos\theta\,\hat{\boldsymbol{x}} \;-\; I\left(k^2 + \frac{ik}{r} - \frac{1}{r^2}\right)\frac{\mathrm{e}^{ikr}}{4\pi r}\,\sin\theta\,\hat{\boldsymbol{\theta}} \tag{2.47b}$$

In the special case $\sigma = 0$ and $\boldsymbol{z} = \boldsymbol{o}$ the far field patterns are given by

$$\boldsymbol{E}_\infty(\theta,\phi) \;=\; -\frac{i\omega\mu_0 I}{4\pi}\,\sin\theta\,\hat{\boldsymbol{\theta}}\,, \qquad \boldsymbol{H}_\infty(\theta,\phi) \;=\; -\frac{i\omega\mu_0 I}{4\pi}\,\sin\theta\,\hat{\boldsymbol{\phi}} \tag{2.48a}$$

for the electric dipole and

$$\boldsymbol{E}_\infty(\theta,\phi) \;=\; \frac{k\omega\mu_0 M}{4\pi}\,\sin\theta\,\hat{\boldsymbol{\theta}}\,, \qquad \boldsymbol{H}_\infty(\theta,\phi) \;=\; -\frac{k\omega\mu_0 M}{4\pi}\,\sin\theta\,\hat{\boldsymbol{\phi}} \tag{2.48b}$$

for the magnetic dipole. The radiated power from (2.35) takes the form

$$P_\infty \;=\; \frac{\omega^2\mu_0^2 I^2}{12\pi}\,Y_0,\,, \qquad P_\infty \;=\; \frac{k^2\omega^2\mu_0^2 M^2}{12\pi}\,Y_0\,, \tag{2.49}$$

respectively.

We would like now to return to our example in Section 1.2. There we introduced the notion of an array by assuming a (finite) number of electric dipoles at locations $\boldsymbol{y}_n$, $n = -N, \ldots, N$. Let $a_n\,\hat{\boldsymbol{p}}$ be the common dipole moment. Then, according to (2.44b), the n^{th} dipole generates the electric far field pattern

$$\boldsymbol{E}_{n,\infty}(\hat{\boldsymbol{x}}) \;=\; a_n\,\frac{i\omega\mu_0}{4\pi}\,\mathrm{e}^{-ik\boldsymbol{y}_n\cdot\hat{\boldsymbol{x}}}\,\left[\hat{\boldsymbol{x}} \times \left(\hat{\boldsymbol{p}} \times \hat{\boldsymbol{x}}\right)\right].$$

The whole array generates the far field pattern by superposition, i.e.

$$\boldsymbol{E}_\infty(\hat{\boldsymbol{x}}) \;=\; \frac{i\omega\mu_0}{4\pi}\,\hat{\boldsymbol{x}} \times \left(\hat{\boldsymbol{p}} \times \hat{\boldsymbol{x}}\right)\,\sum_{n=-N}^{N} a_n\,\mathrm{e}^{-ik\boldsymbol{y}_n\cdot\hat{\boldsymbol{x}}}\,.$$

This formula coincides with (1.3).

The electromagnetic fields of a *finite line current* flowing along the straight line $\boldsymbol{x} = s\,\hat{\boldsymbol{e}}_3$, $s \in [-\ell, \ell]$, of length 2ℓ and direction $\hat{\boldsymbol{e}}_3$ can be modeled by the limiting process of an array, when the distance d between the elements tends to zeros and the number of elements to infinity. This leads to the determination of $\boldsymbol{H}$ and $\boldsymbol{E}$ from the potential

$$u(\boldsymbol{x}) \;=\; \int_{-\ell}^{\ell} I(s)\,\Phi(\boldsymbol{x}, s\hat{\boldsymbol{e}}_3)\,ds \;=\; \int_{-\ell}^{\ell} I(s)\,\frac{\exp(ik\,|\boldsymbol{x} - s\hat{\boldsymbol{e}}_3|)}{4\pi\,|\boldsymbol{x} - s\hat{\boldsymbol{e}}_3|}\,ds \tag{2.50}$$

via

$$\boldsymbol{H}(\boldsymbol{x}) \;=\; \operatorname{curl}\left(u(\boldsymbol{x})\,\hat{\boldsymbol{e}}_3\right), \qquad \boldsymbol{E}(\boldsymbol{x}) \;=\; i\omega\mu_0\, u(\boldsymbol{x})\,\hat{\boldsymbol{e}}_3 + \frac{1}{\sigma - i\omega\epsilon}\,\nabla\big(\partial u(\boldsymbol{x})/\partial x_3\big)\,. \tag{2.51}$$

The far field patterns are computed, using (2.44a)–(2.44d), by

$$\begin{aligned}\boldsymbol{E}_\infty(\hat{\boldsymbol{x}}) &= \frac{i\omega\mu_0}{4\pi}\left[\hat{\boldsymbol{x}} \times (\hat{\boldsymbol{e}}_3 \times \hat{\boldsymbol{x}})\right] \int\limits_{-\ell}^{\ell} I(s)\,\mathrm{e}^{-iks\hat{\boldsymbol{x}}\cdot\hat{\boldsymbol{e}}_3}\,ds \\ &= \frac{i\omega\mu_0}{4\pi}\left[\hat{\boldsymbol{x}} \times (\hat{\boldsymbol{e}}_3 \times \hat{\boldsymbol{x}})\right] \int\limits_{-\ell}^{\ell} I(s)\,\mathrm{e}^{-iks\cos\theta}\,ds\,, \qquad (2.52\text{a})\\ \boldsymbol{H}_\infty(\hat{\boldsymbol{x}}) &= \frac{ik}{4\pi}\,(\hat{\boldsymbol{x}} \times \hat{\boldsymbol{e}}_3) \int\limits_{-\ell}^{\ell} I(s)\,\mathrm{e}^{-iks\hat{\boldsymbol{x}}\cdot\hat{\boldsymbol{e}}_3}\,ds \\ &= \frac{ik}{4\pi}\,(\hat{\boldsymbol{x}} \times \hat{\boldsymbol{e}}_3) \int\limits_{-\ell}^{\ell} I(s)\,\mathrm{e}^{-iks\cos\theta}\,ds\,. \qquad (2.52\text{b})\end{aligned}$$

2.11 Boundary Conditions on Interfaces

If we consider a situation in which a surface S separates two homogeneous media from each other, the constitutive parameters ϵ, μ and σ are no longer continuous but piecewise continuous with finite jumps on S. While on both sides of S Maxwell's equations (2.5a)–(2.5d) hold, the presence of these jumps implies that the fields satisfy certain conditions on the surface.

To derive the mathematical form of this behaviour (the boundary conditions) we apply the law of induction (2.1b) to a narrow rectangle-like surface C, containing the normal n to the surface S and whose long sides C_+ and C_- are parallel to S and are on the opposite sides of it, cf. Figure 2.1.

When we let the height of the narrow sides, AA' and BB', approach zero C_+ and C_- approach a curve C on S, the surface integral $\frac{\partial}{\partial t}\int_R \boldsymbol{\mathcal{B}} \cdot \boldsymbol{\nu}\, dS$ will vanish in the limit since the field remains finite (note, that the normal $\boldsymbol{\nu}$ is the normal to R lying in the tangential plane of S). Hence, the line integrals $\int_C \boldsymbol{\mathcal{E}}_+ \cdot \boldsymbol{t}\, d\ell$ and $\int_C \boldsymbol{\mathcal{E}}_- \cdot \boldsymbol{t}\, d\ell$ must be equal. Since the curve C is arbitrary the integrands $\boldsymbol{\mathcal{E}}_+ \cdot \boldsymbol{t}$ and $\boldsymbol{\mathcal{E}}_- \cdot \boldsymbol{t}$ coincide on every arc C, i.e.

$$\boldsymbol{n} \times \boldsymbol{\mathcal{E}}_+ \;-\; \boldsymbol{n} \times \boldsymbol{\mathcal{E}}_- \;=\; \boldsymbol{o} \quad \text{on } S. \tag{2.53}$$

A similar argument holds for the magnetic field in (2.1a) if the current distribution $\boldsymbol{\mathcal{J}} = \sigma\boldsymbol{\mathcal{E}} + \boldsymbol{\mathcal{J}}_e$ remains finite. In this case, the same arguments lead to the boundary condition

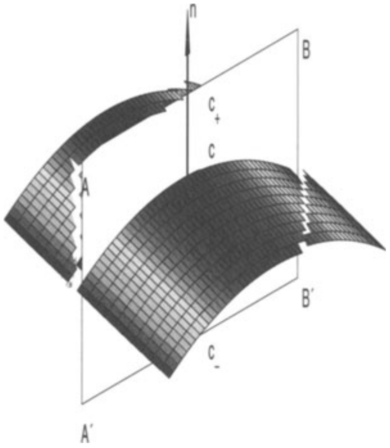


Fig. 2.1. The derivation of the boundary conditions

$$\boldsymbol{n} \times \boldsymbol{\mathcal{H}}_+ \; - \; \boldsymbol{n} \times \boldsymbol{\mathcal{H}}_- \; = \; \boldsymbol{o} \quad \text{on } S. \tag{2.54}$$

If, however, the external current distribution is a surface current, i.e. if $\boldsymbol{\mathcal{J}}_e$ is of the form $\boldsymbol{\mathcal{J}}_e(\boldsymbol{x} + \tau \boldsymbol{n}(\boldsymbol{x})) = \boldsymbol{\mathcal{J}}_s(\boldsymbol{x})\delta(\tau)$ for small τ and $\boldsymbol{x} \in S$ and with tangential surface field $\boldsymbol{\mathcal{J}}_s$ and σ is finite, then the surface integral $\int_R \boldsymbol{\mathcal{J}}_e \cdot \boldsymbol{\nu}\, dS$ will tend to $\int_C \boldsymbol{\mathcal{J}}_s \cdot \boldsymbol{\nu}\, d\ell$, and so the boundary condition is

$$\boldsymbol{n} \times \boldsymbol{\mathcal{H}}_+ \; - \; \boldsymbol{n} \times \boldsymbol{\mathcal{H}}_- \; = \; \boldsymbol{\mathcal{J}}_s \quad \text{on } S. \tag{2.55}$$

We will call (2.53) and (2.54) or (2.55) the **transmission boundary conditions**.

In many applications it is also important to consider the case in which the interface S is covered by a thin layer of very high conductivity, i.e. $\sigma(\boldsymbol{x} + \tau \boldsymbol{n}(x)) = \sigma_s(\boldsymbol{x})\delta(\tau)$ for small τ and $\boldsymbol{x} \in S$ and with surface conductivity σ_s. If $\boldsymbol{\mathcal{J}}_e$ remains finite then the surface integral $\int_R \boldsymbol{\mathcal{J}} \cdot \boldsymbol{n}\, dS$ will tend to $\int_C \sigma_s \boldsymbol{\mathcal{E}} \cdot \boldsymbol{n}\, d\ell + \int_C \boldsymbol{\mathcal{J}}_s \cdot \boldsymbol{n}\, d\ell$, i.e.

$$\boldsymbol{n} \times \boldsymbol{\mathcal{H}}_+ \; - \; \boldsymbol{n} \times \boldsymbol{\mathcal{H}}_- \; = \; \sigma_s \boldsymbol{n} \times (\boldsymbol{\mathcal{E}} \times \boldsymbol{n}) \; + \; \boldsymbol{\mathcal{J}}_s \quad \text{on } S. \tag{2.56}$$

We will call this the **conductive boundary condition**. This condition has been used (see e.g. [121],[122]) to model the situation in which the field penetrates the object only to a small depth. Thus this condition is closely related to the transmission conditions as well as to the impedance, or Leontovich, condition which we mention below.

A special and very important case is that of a **perfectly conducting medium** with boundary S. Such a medium is characterized by the fact that the electric field vanishes inside this medium, and (2.53) reduces to

$$\boldsymbol{n} \times \boldsymbol{\mathcal{E}} = \boldsymbol{o} \quad \text{on } S \tag{2.57}$$

Another important case is the **impedance-** or **Leontovich boundary condition**

$$\boldsymbol{n} \times \boldsymbol{\mathcal{H}} = \lambda \boldsymbol{n} \times (\boldsymbol{\mathcal{E}} \times \boldsymbol{n}) \quad \text{on } S \tag{2.58}$$

which, under appropriate conditions, may be used as an approximation of the transmission conditions [120].

Finally, we specify the boundary conditions to the E- and H-modes derived Section 2.8. We assume that the surface S is an infinite cylinder in x_3–direction with constant cross section. Furthermore, we assume that the volume current density j vanishes near the boundary S and that the surface current densities take the form $\boldsymbol{J}_s = j_s \hat{e}_3$ for the E-mode and $\boldsymbol{J}_s = j_s(\boldsymbol{n} \times \hat{e}_3)$ for the H-mode. We use the notation $[v] := v|_+ - v|_-$ for the jump of the function v at the boundary. Also, we abbreviate (only for this table) $\sigma' = \sigma - i\omega\epsilon$. We list the boundary conditions in the following table.

Bound. cond.	E-mode	H-mode
transmission	$[k^2 u] = 0$ on S, $[\sigma' \frac{\partial u}{\partial n}] = -j_s$ on S,	$[\mu \frac{\partial u}{\partial n}] = 0$ on S, $[k^2 u] = j_s$ on S,
conductive	$[k^2 u] = 0$ on S, $-[\sigma' \frac{\partial u}{\partial n}] = \sigma_s k^2 u + j_s$,	$[\mu \frac{\partial u}{\partial n}] = 0$ on S, $[k^2 u] = \sigma_s i\omega\mu \frac{\partial u}{\partial n} + j_s$,
impedance	$\lambda k^2 u + \sigma' \frac{\partial u}{\partial n} = -j_s$ on S,	$k^2 u - \lambda i\omega\mu \frac{\partial u}{\partial n} = j_s$ on S,
perfect conductor	$u = 0$ on S,	$\frac{\partial u}{\partial n} = 0$ on S.

2.12 Hertz Potentials and Classes of Solutions

In this section we recall some of the most important classes of solutions of Maxwell's equations (2.13a)–(2.13d) in homogeneous, isotropic and source free media. We use the constructions with the Hertz potentials in the TM and TE modes described in Section 2.8.

(A) Plane waves:

First, we take $\boldsymbol{A}_e = \boldsymbol{o}$, $\boldsymbol{A}_m(\boldsymbol{x}) = -1/(k\omega\mu)\exp(ik\hat{\boldsymbol{\alpha}}\cdot\boldsymbol{x})\,\boldsymbol{a}$ for some fixed vector $\boldsymbol{a} \in \mathbb{C}^3$ and unit vector $\hat{\boldsymbol{\alpha}} \in \mathbb{R}^3$. This results in **plane waves:**

$$\boldsymbol{E}(\boldsymbol{x}) = (\hat{\boldsymbol{\alpha}} \times \boldsymbol{a})\,\mathrm{e}^{ik\hat{\boldsymbol{\alpha}}\cdot\boldsymbol{x}}\,, \tag{2.59a}$$

$$\boldsymbol{H}(\boldsymbol{x}) = \frac{k}{\omega\mu}\,\hat{\boldsymbol{\alpha}} \times (\hat{\boldsymbol{\alpha}} \times \boldsymbol{a})\,\mathrm{e}^{ik\hat{\boldsymbol{\alpha}}\cdot\boldsymbol{x}}\,. \tag{2.59b}$$

The corresponding time dependent waves are

$$\boldsymbol{\mathcal{E}}(\boldsymbol{x},t) = (\hat{\boldsymbol{\alpha}} \times \boldsymbol{a})\,\mathrm{e}^{-\operatorname{Im}k\,\hat{\boldsymbol{\alpha}}\cdot\boldsymbol{x}}\,\mathrm{e}^{i\operatorname{Re}k\,\hat{\boldsymbol{\alpha}}\cdot\boldsymbol{x}-i\omega t}\,,$$

$$\boldsymbol{\mathcal{H}}(\boldsymbol{x},t) = \frac{k}{\omega\mu}\,\hat{\boldsymbol{\alpha}} \times (\hat{\boldsymbol{\alpha}} \times \boldsymbol{a})\,\mathrm{e}^{-\operatorname{Im}k\,\hat{\boldsymbol{\alpha}}\cdot\boldsymbol{x}}\,\mathrm{e}^{i\operatorname{Re}k\,\hat{\boldsymbol{\alpha}}\cdot\boldsymbol{x}-i\omega t}\,.$$

The phase factor is constant on the planes $\operatorname{Re}k\,\hat{\boldsymbol{\alpha}}\cdot\boldsymbol{x} = \omega t + \delta$ traveling with velocity $v = \omega/\operatorname{Re}k$ (**phase velocity**) in the direction $\hat{\boldsymbol{\alpha}}$. For non-conductive media $v = 1/\sqrt{\mu\epsilon}$. We see that $\boldsymbol{H} = \frac{k}{\omega\mu}\,\hat{\boldsymbol{\alpha}} \times \boldsymbol{E} = Y_0\,\hat{\boldsymbol{\alpha}} \times \boldsymbol{E}$. The quantity

$$Y_0 \;:=\; \frac{k}{\omega\mu} \tag{2.60}$$

is called the **intrinsic admittance** of the medium which is equal to $\sqrt{\epsilon/\mu}$ for non-conductive media (see (2.30)).

(B) Spherical waves:

As a second class of solutions we take $\boldsymbol{A}_e = \boldsymbol{o}$ and $\boldsymbol{A}_m(\boldsymbol{x}) = Y_n^m(\theta,\phi)\,\lambda_n(kr)\,r\,\hat{\boldsymbol{x}}$ where r,θ,ϕ are the spherical polar coordinates of x and $\hat{\boldsymbol{x}}$ denotes the coordinate vector in r–direction. $Y_n^m(\theta,\phi) = P_n^m(\cos\theta)\exp(im\phi)$, $|m| \leq n$, $n \in \mathbb{N}_0$, denote the spherical harmonics where we have denoted the associated Legendre function of order n and degree m by Y_n^m. By λ_n we denote either the **spherical Bessel function** j_n or the **spherical Hankel functions** $h_n^{(1)}$ or $h_n^{(2)}$ of the first and second kind, respectively, and order of n. We refer to [139, 50, 30] for an introduction into Bessel- and Hankel functions. Since $Y_n^m(\theta,\phi)\,\lambda_n(kr)$ are solutions of the Helmholtz equation (2.21) for $r > 0$ it is easily seen that $\boldsymbol{A}_m$ satisfies the inhomogeneous vector Helmholtz equation

$$\Delta\boldsymbol{A}_m \;+\; k^2\boldsymbol{A}_m \;=\; 2\nabla\big[Y_n^m(\theta,\phi)\,\lambda_n(kr)\big] \quad \text{for } r > 0.$$

The fields

$$\boldsymbol{E}(\boldsymbol{x}) = i\omega\,\mu\,\mathrm{curl}\,\left[Y_n^m(\theta,\phi)\,\lambda_n(kr)\,r\,\hat{\boldsymbol{x}}\right] \tag{2.61a}$$

$$= -i\omega\,\mu\,\hat{\boldsymbol{x}} \times \nabla\left[Y_n^m(\theta,\phi)\,\lambda_n(kr)\,r\right] \tag{2.61b}$$

$$\boldsymbol{H}(\boldsymbol{x}) = \frac{1}{i\omega\,\mu}\mathrm{curl}\,\boldsymbol{E}(\boldsymbol{x}) \;=\; \mathrm{curl}^2\left[Y_n^m(\theta,\phi)\,\lambda_n(kr)\,r\,\hat{\boldsymbol{x}}\right] \tag{2.61c}$$

are called **toroidal fields**. Analogously, $\boldsymbol{A}_m = \boldsymbol{o}$ and $\boldsymbol{A}_e(\boldsymbol{x}) = Y_n^m(\theta,\phi)\,\lambda_n(kr)\,r\,\hat{\boldsymbol{x}}$ lead to **spheroidal fields** of the form

$$\boldsymbol{H}(\boldsymbol{x}) = (\sigma - i\omega\,\epsilon)\,\mathrm{curl}\,\left[Y_n^m(\theta,\phi)\,\lambda_n(kr)\,r\,\hat{\boldsymbol{x}}\right], \tag{2.62a}$$

$$\boldsymbol{E}(\boldsymbol{x}) = \frac{1}{\sigma - i\omega\,\epsilon}\,\mathrm{curl}\,\boldsymbol{H}(\boldsymbol{x}) \;=\; \mathrm{curl}^{\,2}\left[Y_n^m(\theta,\phi)\,\lambda_n(kr)\,r\,\hat{\boldsymbol{x}}\right]. \tag{2.62b}$$

The fields with $\lambda_n = j_n$ are smooth at the origin while the fields with $\lambda_n = h_n^{(1),(2)}$ are singular at the origin.

(C) Cylindrical waves:

Electromagnetic waves in waveguides are described by using cylindrical coordinates ρ, ϕ, z. We set $\boldsymbol{A}_e = \boldsymbol{o}$ and $\boldsymbol{A}_m(x) = \boldsymbol{a}\,\exp(i\beta z)\,\Lambda_n(\kappa\rho)\,\exp(in\phi)$ with constant vector $\boldsymbol{a} \in \mathbb{C}^3$ and $\beta \in \mathbb{R}$ where $\kappa = \sqrt{k^2 - \beta^2}$ (with $\mathrm{Re}\,\kappa \geq 0$ and $\mathrm{Im}\,\kappa \geq 0$). Here, Λ_n denotes one of the **cylindrical Bessel functions** J_n or **Hankel functions** $H_n^{(1)}$, $H_n^{(2)}$ of the first and second kind, respectively, and of order n. Then $\Lambda_n(\kappa\rho)\,\exp(in\phi)$ solves the two dimensional Helmholtz equation $\Delta u + \kappa^2 u = 0$ for $\rho > 0$. We arrive at the fields

$$\boldsymbol{E}(\boldsymbol{x}) = -i\omega\,\mu\,\boldsymbol{a} \times \nabla\left[\exp(i\beta z)\,\Lambda_n(\kappa\rho)\,\exp(in\phi)\right], \tag{2.63a}$$

$$\boldsymbol{H}(\boldsymbol{x}) = \frac{1}{i\omega\,\mu}\mathrm{curl}\,\boldsymbol{E}(\boldsymbol{x}) \;=\; \mathrm{curl}^{\,2}\left[\exp(i\beta z)\,\Lambda_n(\kappa\rho)\,\exp(in\phi)\,\boldsymbol{a}\right] \tag{2.63b}$$

which are smooth at the line $\rho = 0$ only if $\Lambda_n = J_n$.

Now we check which of these special solutions of Maxwell's equations satisfy the radiation conditions (2.28) or (2.29a), (2.29b). We restrict ourselves to the case of k being real.

From (2.31a), (2.31b) we conclude that a necessary (but not sufficient!) condition for the Silver-Müller radiation condition (2.29a), (2.29b) to hold is that the fields decay as $1/r$ when r tends to infinity. From this we see that no plane or cylindrical wave satisfies the Silver-Müller radiation conditions (for the latter see Section 2.13!).

From the asymptotic behaviour of the spherical Hankel functions

$$h_n^{(1),(2)}(t) = \frac{1}{t}\,\mathrm{e}^{\pm i\left(t-(n+1)\pi/2\right)}\left\{1 \,+\, \mathcal{O}(1/t)\right\} \quad \text{as } t \to \infty \tag{2.64a}$$

$$\frac{d}{dt}h_n^{(1),(2)}(t) = \frac{1}{t}\,\mathrm{e}^{\pm i\left(t-n\pi/2\right)}\left\{1 \,+\, \mathcal{O}(1/t)\right\} \quad \text{as } t \to \infty \tag{2.64b}$$

$$j_n(t) = \frac{1}{t}\,\cos\bigl(t-(n+1)\pi/2\bigr)\left\{1 \,+\, \mathcal{O}(1/t)\right\} \quad \text{as } t \to \infty \tag{2.64c}$$

$$\frac{d}{dt}j_n(t) = -\frac{1}{t}\,\sin\bigl(t-(n+1)\pi/2\bigr)\left\{1 \,+\, \mathcal{O}(1/t)\right\} \quad \text{as } t \to \infty. \tag{2.64d}$$

we conclude that only the spherical wave functions

$$\boldsymbol{E}(\boldsymbol{x}) = i\omega\,\mu\,\mathrm{curl}\,\left[Y_n^m(\theta,\phi)\,h_n^{(1)}(kr)\,r\,\hat{\boldsymbol{x}}\right]$$

$$\boldsymbol{H}(\boldsymbol{x}) = \frac{1}{i\omega\,\mu}\mathrm{curl}\,\boldsymbol{E}(\boldsymbol{x}) \;=\; \mathrm{curl}^{\,2}\left[Y_n^m(\theta,\phi)\,h_n^{(1)}(kr)\,r\,\hat{\boldsymbol{x}}\right]$$

and

$$\begin{aligned} \boldsymbol{H}(\boldsymbol{x}) &= (\sigma - i\omega\,\epsilon)\,\mathrm{curl}\,\Big[Y_n^m(\theta,\phi)\,h_n^{(1)}(kr)\,r\,\hat{\boldsymbol{x}}\Big]\,, \\ \boldsymbol{E}(\boldsymbol{x}) &= \frac{1}{\sigma - i\omega\,\epsilon}\,\mathrm{curl}\,\boldsymbol{H}(\boldsymbol{x}) \;\doteq\; \mathrm{curl}^{\,2}\Big[Y_n^m(\theta,\phi)\,h_n^{(1)}(kr)\,r\,\hat{\boldsymbol{x}}\Big] \end{aligned}$$

satisfy the Silver-Müller radiation condition with far field patterns

$$\boldsymbol{E}_\infty(\theta,\phi) = \frac{i\omega\,\mu}{k}\,\mathrm{e}^{-\pi(n+1)i/2}\left[\frac{\partial}{\partial\theta}Y_n^m(\theta,\phi)\,\hat{\boldsymbol{\phi}} - \frac{1}{\sin\theta}\,\frac{\partial}{\partial\phi}Y_n^m(\theta,\phi)\,\hat{\boldsymbol{\theta}}\right] \tag{2.65a}$$

$$\boldsymbol{H}_\infty(\theta,\phi) = \hat{\boldsymbol{x}} \times \boldsymbol{E}_\infty(\theta,\phi), \tag{2.65b}$$

and

$$\begin{aligned} \boldsymbol{H}_\infty(\theta,\phi) &= \frac{\sigma - i\omega\,\epsilon}{k}\,\mathrm{e}^{-\pi(n+1)i/2}\left[\frac{\partial}{\partial\theta}Y_n^m(\theta,\phi)\,\hat{\boldsymbol{\phi}} - \frac{1}{\sin\theta}\,\frac{\partial}{\partial\phi}Y_n^m(\theta,\phi)\,\hat{\boldsymbol{\theta}}\right] \\ \boldsymbol{E}_\infty(\theta,\phi) &= \boldsymbol{H}_\infty(\theta,\phi) \times \hat{\boldsymbol{x}}\,, \end{aligned}$$

respectively.

2.13 Radiation Problems in Two Dimensions

We have seen in the previous sections that the outgoing spherical waves and the fields generated by electric and magnetic dipoles satisfy the Silver-Müller radiation condition (2.29a), (2.29b) if k is real. As we mentioned above, this radiation condition describes the correct behaviour of the radiating fields generated by sources lying in a compact set. If, however, the sources are distributed on an unbounded set e.g., along an infinite line, the fields decay more slowly. First we look at the cylindrical waves again for the special case where k is real and positive, $\beta = 0$, $\boldsymbol{a} = \hat{\boldsymbol{e}}_3$ and $\Lambda_n = H_n^{(1)}$. With cylindrical coordinates ρ, ϕ, z we now have

$$\boldsymbol{E}(\boldsymbol{x}) = -i\omega\,\mu\,\hat{\boldsymbol{e}}_3 \times \nabla\Big[H_n^{(1)}(\kappa\rho)\,\exp(in\phi)\Big]\,, \tag{2.66a}$$

$$\boldsymbol{H}(\boldsymbol{x}) = \frac{1}{i\omega\,\mu}\mathrm{curl}\,\boldsymbol{E}(\boldsymbol{x}) \;=\; \mathrm{curl}^{\,2}\Big[H_n^{(1)}(\kappa\rho)\,\exp(in\phi)\,\hat{\boldsymbol{e}}_3\Big] \tag{2.66b}$$

We have seen that $\boldsymbol{E}$ and $\boldsymbol{H}$ do not satisfy the Silver-Müller radiation condition (2.29a), (2.29b) since they are only bounded in x_3–direction but do not decay to zero. We will now show that they satisfy a weaker form of the radiation condition in the (x_1, x_2)–plane. Let us set $u(\rho,\phi) = H_n^{(1)}(\kappa\rho)\,\exp(in\phi)$. Then u satisfies the **two dimensional Helmholtz equation**

$$\Delta u \;+\; \kappa^2\,u \;=\; 0 \quad \text{in } \mathbb{R}^2 \setminus \{o\}\,. \tag{2.67}$$

Furthermore, from the asymptotic behaviour of the Hankel functions

$$H_n^{(1)}(t) = \sqrt{\frac{2}{\pi t}}\, \mathrm{e}^{i\left(t-n\pi/2-\pi/4\right)} \left\{1 \;+\; \mathcal{O}(1/t)\right\} \quad \text{as } t \to \infty \tag{2.68a}$$

$$\frac{d}{dt} H_n^{(1)}(t) = \sqrt{\frac{2}{\pi t}}\, \mathrm{e}^{i\left(t-n\pi/2+\pi/4\right)} \left\{1 \;+\; \mathcal{O}(1/t)\right\} \quad \text{as } t \to \infty \tag{2.68b}$$

we conclude that u satisfies the **two dimensional Sommerfeld radiation condition**

$$\frac{\partial}{\partial r} u(x) \;-\; i\,\kappa\, u(x) \;=\; \mathcal{O}\big(1/\rho^{3/2}\big) \quad \text{as } \rho \to \infty \tag{2.69}$$

uniformly with respect to $\phi \in [0, 2\pi]$.

Let u be any solution u of the Helmholtz equation (2.67) satisfying the radiation condition (2.69). Then it can be shown (see [29]) that u and ∇u decay as $\mathcal{O}(1/\sqrt{\rho})$. Furthermore, for

$$\begin{aligned} \boldsymbol{E}(\boldsymbol{x}) &= -i\omega\,\mu\,\hat{\boldsymbol{e}}_3 \times \nabla u(\rho,\phi) \\ &= i\omega\,\mu \left[\frac{1}{\rho}\,\frac{\partial u(\rho,\phi)}{\partial \phi}\,\hat{\boldsymbol{\rho}} \;-\; \frac{\partial u(\rho,\phi)}{\partial \rho}\,\hat{\boldsymbol{\phi}}\right] \end{aligned} \tag{2.70a}$$

$$\begin{aligned} \boldsymbol{H}(\boldsymbol{x}) &= \frac{1}{i\omega\,\mu} \operatorname{curl} \boldsymbol{E}(\boldsymbol{x}) \;=\; \operatorname{curl}^2\big(u(\rho,\phi)\,\hat{\boldsymbol{e}}_3\big) \\ &= -\Delta u(\rho,\phi)\,\hat{\boldsymbol{e}}_3 \;=\; \kappa^2\, u(\rho,\phi)\,\hat{\boldsymbol{e}}_3 \end{aligned} \tag{2.70b}$$

since $\operatorname{div}(\hat{\boldsymbol{e}}_3 u) = 0$. From this we conclude that the cylindrical waves $\boldsymbol{E}$ and $\boldsymbol{H}$ from (2.70a), (2.70b) satisfy the **cylindrical Silver-Müller radiation condition**

$$\boldsymbol{E}(\boldsymbol{x}) \times \hat{\boldsymbol{\rho}} \;+\; \frac{1}{Y_0}\,\boldsymbol{H}(\boldsymbol{x}) = \mathcal{O}\big(1/\rho^{3/2}\big)\,, \quad \rho \to \infty\,, \tag{2.71a}$$

$$\boldsymbol{H}(\boldsymbol{x}) \times \hat{\boldsymbol{\rho}} \;-\; Y_0\,\boldsymbol{E}(\boldsymbol{x}) = \mathcal{O}\big(1/\rho^{3/2}\big)\,, \quad \rho \to \infty\,, \tag{2.71b}$$

uniformly with respect to $\phi \in [0, \pi]$. Again, $Y_0 = \frac{k}{\omega\mu}$ denotes the admittance from (2.30).

The Sommerfeld radiation condition (2.69) implies that u and ∇u have the asymptotic forms

$$u(\rho,\phi) = \frac{\exp(i\kappa\rho)}{\sqrt{\rho}}\, u_\infty(\phi) \;+\; \mathcal{O}(1/\rho^{3/2})\,, \quad \rho \to \infty\,, \tag{2.72a}$$

$$\begin{aligned} \frac{\partial u(\rho,\phi)}{\partial \rho} &= \frac{\partial}{\partial \rho}\frac{\exp(i\kappa\rho)}{\sqrt{\rho}}\, u_\infty(\phi) \;+\; \mathcal{O}(1/\rho^{3/2})\,, \quad \rho \to \infty\,, \\ &= i\kappa\, \frac{\exp(i\kappa\rho)}{\sqrt{\rho}}\, u_\infty(\phi) \;+\; \mathcal{O}(1/\rho^{3/2})\,, \quad \rho \to \infty\,, \end{aligned} \tag{2.72b}$$

$$\frac{\partial u(\rho,\phi)}{\partial \phi} = \frac{\exp(i\kappa\rho)}{\sqrt{\rho}}\, \frac{d u_\infty(\phi)}{d\phi} \;+\; \mathcal{O}(1/\rho^{3/2})\,, \quad \rho \to \infty\,, \tag{2.72c}$$

uniformly with respect to ϕ. Again we call the non-radial function u_∞ the **far field pattern** of the scalar potential u. This asymptotic form of u yields immediately the corresponding asymptotic behaviour

$$\boldsymbol{E}(\boldsymbol{x}) = \omega\mu\kappa \frac{\exp(i\kappa\rho)}{\sqrt{\rho}} u_\infty(\phi)\, \hat{\boldsymbol{\phi}} \;+\; \mathcal{O}(1/\rho^{3/2}), \quad \rho \to \infty, \tag{2.73a}$$

$$\boldsymbol{H}(\boldsymbol{x}) = \kappa^2 \frac{\exp(i\kappa\rho)}{\sqrt{\rho}} u_\infty(\phi)\, \hat{\boldsymbol{e}}_3 \;+\; \mathcal{O}(1/\rho^{3/2}), \quad \rho \to \infty, \tag{2.73b}$$

The vector fields

$$\boldsymbol{E}_\infty(\phi) \;:=\; \omega\mu\kappa\, u_\infty(\phi)\, \hat{\boldsymbol{\phi}} \quad \text{and} \quad \boldsymbol{H}_\infty(\phi) \;:=\; \omega\mu\kappa\, u_\infty(\phi)\, \hat{\boldsymbol{e}}_3$$

are called the **far field pattern** of the two dimensional vector fields $\boldsymbol{E}$ and $\boldsymbol{H}$. They satisfy also

$$\boldsymbol{E}(\boldsymbol{x}) = \frac{\exp(i\kappa\rho)}{\sqrt{\rho}} \boldsymbol{E}_\infty(\phi) \;+\; \mathcal{O}(1/\rho^{3/2}), \quad \rho \to \infty, \tag{2.74a}$$

$$\boldsymbol{H}(\boldsymbol{x}) = Y_0 \frac{\exp(i\kappa\rho)}{\sqrt{\rho}} \boldsymbol{H}_\infty(\phi) \;+\; \mathcal{O}(1/\rho^{3/2}), \quad \rho \to \infty, \tag{2.74b}$$

and

$$\boldsymbol{H}_\infty(\phi) \;=\; \hat{\boldsymbol{\rho}} \times \boldsymbol{E}_\infty(\phi), \quad \hat{\boldsymbol{\rho}} \cdot \boldsymbol{E}_\infty(\phi) \;=\; \hat{\boldsymbol{\rho}} \cdot \boldsymbol{H}_\infty(\phi) \;=\; 0 \quad \text{for all } \phi, \tag{2.75}$$

compare with (2.32).

3

Optimization Theory for Antennas

3.1 Introductory Remarks

Considering the variety of physical forms antennas may take, it is understandable, that the analysis of such electromagnetic devices traditionally proceeds by discussion of individual cases. We gave a preliminary discussion for arrays and line sources in Chapter 1. If one refers back to that chapter while reading the material of this one, he will see there the basic ideas that we treat systematically here in Chapter 3. We will take a general approach in order to construct a single setting in which we may treat a variety of situations including arrays and line sources.

Regardless of the specific physical structure, mathematical questions arise from the desire to control and even optimize antenna performance through the manipulation of certain parameters. In addition to the obvious choice of "feeding", these parameters may include measures of structural properties as well. For example, in the case of arrays of elementary radiators, we have seen that these quantities may include, in addition to the feeding, parameters involving the number of elements, the general geometric form (e.g. linear, or circular), and the spacings between the radiators. Each concrete choice of antenna type determines the set of parameters which may be varied, although common to all is the idea of manipulating currents in or on the structure as a means of controlling the radiation characteristics.

Optimization of course, implies some *a priori* measure of performance. Such performance criteria most often involve some numerical functional which depends explicitly on the asymptotic form of the electromagnetic field i.e., on the far field radiation pattern or on a related quantity.[1] Such was the case in the discussion of the directivity of arrays in Chapter 1. Classical examples of performance measures found frequently in the literature are directivity, gain,

[1] Important problems arise concerning the optimization of near-field quantities which may be treated by these methods.

signal-to-noise ratio, and various "quality factors". In Section 3.3 we will discuss specific forms typically associated with different antenna types and configurations. But we remark here for the sake of emphasis, and we mention it later, that there is a significant question concerning choice of definitions. The problem is most easily illustrated in terms of a concrete example.

For simplicity, consider a model of a simple dipole oscillator and its feeding mechanism. One may imagine a generator supplying an alternating current to the antenna establishing a current ψ on the antenna Γ which in turn radiates an electromagnetic field whose far field pattern we denote, as before, by f. In any physical realization, there is a loss of power in the lines between the generator and the antenna Γ. There are apparently two points of view on how to define quantities such as the directivity or the gain. On the one hand, one looks at the system as a whole, and measures the gain in a given direction, $\hat{\boldsymbol{x}}$, by the ratio of power supplied by the generator, denoted by P_G, to the power in the far field in the direction $\hat{\boldsymbol{x}}$ i.e. $|f(\hat{\boldsymbol{x}})|^2/P_G$. Alternatively, one can isolate the antenna itself, consider the current ψ as given, and measure the efficiency as $|f(\hat{\boldsymbol{x}})|^2/\,\|\psi\|^2$. If the latter choice is made, then of course the practical optimization problem has two stages, first the theoretical specification of an optimal choice of ψ and then the practical specification of a method of producing such a current. Both viewpoints occur in the literature.

In this book, we adopt the second point of view, called the **source equation problem** by Rhodes [115], and consider the optimal choice of surface current on the antenna itself as our primary objective. We caution the reader, therefore, that the definitions of performance criteria which involve "input power", e.g. gain or quality factor, will consider the input power as the power in the surface current on the antenna structure itself. This view is certainly consistent with that often taken in discussions of aperture antennas [115]. Practical problems of producing the currents will not be completely ignored, however. We will address such problems by admitting *a priori* constraints which will define classes of *admissible currents* (as well as admissible choice of the other parameters of importance e.g., the spacing between array elements). The optimization problems considered here are thus *constrained* optimization problems.

A moment's thought about our daily experience makes it obvious that we are familiar with the fact that antennas used in radio or television act in two apparently different ways. There is the antenna which broadcasts the signal and the antenna associated with the device which allows us to hear and/or see the transmitted signal. Our approach to the theory concentrates on the *transmitting* mode. We study the problem of finding optimal loadings for the antenna which optimize some characteristic of the radiated far field.

However, the official IEEE definition of an antenna is

> That part of a transmitting or receiving system that is designed to radiate or receive electromagnetic waves.

Indeed, most antennas are considered to be able to act in both modes and, from the simple example with which we began, we can see that how they are treated depends on the particular situation. In either case, certain characteristics of an antenna are the same[2], in particular the radiation pattern and hence such characteristics as directivity or gain.

The fact that the radiation pattern of a given antenna acting in either mode is the same depends on Green's Theorem from which we can derive the Lorentz Reciprocity Theorem for Maxwell's equations. We refer to the book [30] (in particular Section 6.6).

Roughly speaking, the optimization problems addressed in the classical antenna theory literature fall into two categories. The first, which we may call the **synthesis approach**, specifies a desired far field pattern and asks for an admissible current that will produce such a far field. The second involves the choice of some performance criterion, in the form of a real valued functional associated intrinsically with the antenna, and asks for that current which optimizes the chosen functional. The first type has the character of an inverse problem: given a desired output, find that input that will produce that output. Such inverse problems are, in general, not exactly solvable; they are called ill-posed and we will devote Chapter 4 to a discussion of such problems. Regardless of these distinctions, we can bring together all the optimization problems for antennas into a unified model which we describe in Section 3.2. As we will see, the use of constraints for such optimization problems is often crucial for a successful resolution.

Our first goal is to formulate a general class of constrained optimization problems into which, as a group, the particular antenna optimization problems fall. Formally, a constrained problem for minimizing a real valued functional over a set U is written as

$$\text{Minimize} \quad \mathcal{J}(\psi) \quad \text{over } \psi \in U\,. \tag{3.1}$$

When we describe *general* optimization problems in this and subsequent chapters, we denote by ψ the quantity which we can vary in the set U of **admissible** currents in order to optimize the functional. The element ψ as well as the set U depend of the particular antenna model. For example, in the case of an antenna array the set U consists of admissible feeding vectors $\boldsymbol{a} = (a_{-N}, \dots, a_N) \in \mathbb{C}^{2N+1}$ while in the case of a two or three dimensional solid body the set U consists of current distributions, which are scalar or vector fields and thus elements from an infinitely dimensional function space. In principle, we also could think of ψ being the parametrization of the "shape" of the antenna or, for example, constitutive parameters. However, in this book we consider these parameters as being fixed and known. Despite the different nature of ψ in physical situations, we will call ψ the current. We will discuss the structure and typical examples of constraint sets U in Section 3.4 below.

[2] This may not be the case in certain antennas which contain certain *non-reciprocal elements.*

The **performance functional** $\mathcal{J}$ which relates the feeding current, ψ, to the performance index $\mathcal{J}(\psi)$ depends on the particular goal. By choosing different criteria we will be able to cover all of the above mentioned optimization problems. In Section 3.2 we recall some basic facts from general optimization theory in Hilbert spaces, such as existence and uniqueness of optimal solutions, necessary optimality conditions and an approximation result. It is the purpose of this abstract setting to clarify the conditions on the performance functional $\mathcal{J}$ and the set of constraints which are needed for the concrete applications.

Often, $\mathcal{J}$ will be a function of several terms which all depend on the current ψ. The most important one is the far field f which in most of our applications depends linearly on ψ. It does not necessarily have to be a vector field but can as well be a scalar function as in the case of linear antenna array. The dependence of f on ψ is described by a map $\psi \mapsto f$ which implicitly defines the **far field operator** $\mathcal{K}$. In this map $\mathcal{K}$ the shape of the antenna as well as other "constitutive parameters" including such things as wave number, k, and boundary conditions are modeled. It will also be important to study the dependence of the optimal solution and/or the optimal value on such parameters. Even the linear problem with $\mathcal{K}$ depending linearly on ψ presents difficulties since, except in a handful of cases, we have no *explicit* knowledge of the operator $\mathcal{K}$. We may, however deduce some important properties which we will collect in Section 3.3. Some examples, taken mainly from Chapter 1, illustrate the ideas.

Then, in Section 3.4, we introduce some of the most important functionals $\mathcal{J}$, i.e. measures of antenna performance. We will then apply the general results of Section 3.2 by validating the general assumptions concerning the far field operator, $\mathcal{K}$, its domain and range spaces, the constraint set, U, and the performance functional $\mathcal{J}$. In subsequent chapters we will show how this general analysis may be applied in the analysis of specific problems.

3.2 The General Optimization Problem

Keeping in mind that specific examples of functionals and constraints will be given in Section 3.4, we first present a general discussion of the structure of the optimization problem itself whose general form we repeat here for convenience:

$$\text{Minimize (or maximize)} \quad \mathcal{J}(\psi) \quad \text{subject to } \psi \in U\,.$$

To be more specific, let X be a **separable Hilbert space** with norm $\|\cdot\|_X$ and let $U \subset X$ be a subset. In the following, when we write "normed space" or "Hilbert space" we always mean "separable normed space" and "separable Hilbert space", respectively, without always mentioning this explicitly.

The set U will be called the class of **admissible controls**. We seek an element $\psi^o \in U$ for which $\mathcal{J}(\psi^o)$ is an absolute minimum (or maximum) over U. In other words, in the case that we ask for a minimum, we seek a $\psi^o \in U$ satisfying

$$\mathcal{J}(\psi^o) \;\leq\; \mathcal{J}(\psi) \quad \text{for all } \psi \in U\,. \tag{3.2}$$

We note that we can restrict ourselves to minimization problems since maximizing $\mathcal{J}(\psi)$ is equivalent to minimizing $-\mathcal{J}(\psi)$.

3.2.1 Existence and Uniqueness

In this subsection we will study the question of existence and uniqueness of the optimization problem (3.2) under mild assumptions on the functional $\mathcal{J}$ and the set U. The existence results are applications of a general result due to Weierstrass which states that, in any topological space, any sequentially continuous functional attains its maxima and minima on any sequentially compact set. We will apply this result to the Hilbert space X equipped with the weak topology. We recall (see Definition A.51) that a functional $\mathcal{J} : X \longrightarrow \mathbb{R}$ is called **weakly sequentially continuous** provided for every sequence $\{\psi_k\}_{k=1}^{\infty}$ converging weakly to an element $\psi \in X$ we have

$$\lim_{k\to\infty} \mathcal{J}(\psi_k) \;=\; \mathcal{J}(\psi)\,.$$

Analogously, a set $U \subset X$ is called **weakly sequentially compact** if every sequence $\{\psi_k\}_{k=1}^{\infty} \subset U$ contains a weak accumulation point in U. Then we can show a first existence result.

Theorem 3.1. *Assume that the functional $\mathcal{J} : X \longrightarrow \mathbb{R}$ is weakly sequentially continuous on the Hilbert space X and that $U \subset X$ is weakly sequentially compact. Then there exist $\psi^+ \in U$ and $\psi^- \in U$ such that*

$$\mathcal{J}(\psi^-) \;=\; \inf_{\psi\in U} \mathcal{J}(\psi) \quad \textit{and} \quad \mathcal{J}(\psi^+) \;=\; \sup_{\psi\in U} \mathcal{J}(\psi)\,. \tag{3.3}$$

Proof: We restrict ourselves to the case of a minimization problem. By the definition of the infimum we can choose a sequence $\{\psi_k\}_{k=1}^{\infty} \subset U$ such that

$$\lim_{k\to\infty} \mathcal{J}(\psi_k) \;=\; \inf_{\psi\in U} \mathcal{J}(\psi)\,.$$

(At this point we do not exclude the possibility that the infimum is $-\infty$, i.e. we do not assume that $\mathcal{J}$ is bounded below.) The weak compactness of U yields the existence of a weakly convergent subsequence $\psi_{k_j} \rightharpoonup \psi^-$, $j \to \infty$, for some $\psi^- \in U$. From the weak sequentially continuity of $\mathcal{J}$ we conclude that $\mathcal{J}\big(\psi_{k_j}\big) \to \mathcal{J}(\psi^-)$ and thus $\mathcal{J}(\psi^-) = \inf_{\psi\in U} \mathcal{J}(\psi)$. □

As we will see later, some of the most important functionals are *not* weakly continuous. A careful reading of the previous proof shows that this assumption can be relaxed. It is useful to make the following definition:

Definition 3.2. *The functional $\mathcal{J}$ is* **weakly sequentially lower semi-continuous** *on X provided for every sequence $\{\psi_k\}_{k=1}^{\infty}$ converging weakly to an element $\psi \in X$ we have*

$$\liminf_{k \to \infty} \mathcal{J}(\psi_k) \;\geq\; \mathcal{J}(\psi)\,. \tag{3.4}$$

Analogously, we can define the **weakly sequentially upper semi-continuity** *of $\mathcal{J}$.*

Notice that a functional $\mathcal{J}$ is weakly sequentially upper semi-continuous if and only if $-\mathcal{J}$ has the analogous lower semi-continuity property.

Then we have an existence theorem under the following relaxed assumptions.

Theorem 3.3. *Assume that the functional $\mathcal{J} : X \longrightarrow \mathbb{R}$ is weakly sequentially lower semi-continuous and that $U \subset X$ is weakly sequentially compact. Then there exists a $\psi^o \in U$ such that*

$$\mathcal{J}(\psi^o) \;=\; \inf_{\psi \in U} \mathcal{J}(\psi)\,. \tag{3.5}$$

Analogously, the maximum is attained for weakly sequentially upper semi-continuous functions $\mathcal{J}$.

As we will show in Theorem 3.30, many of the specific functionals of interest to us are weakly lower sequentially semi-continuous or even weakly sequentially continuous mappings on X. As a first important example we consider convex functions and begin with a definition.

Definition 3.4. *Let X be a normed space.*

(a) A **subset** *$U \subset X$ is called* **convex** *if*

$$\lambda x + (1-\lambda) y \in U \quad \text{for all } x, y \in U \text{ and } \lambda \in [0,1]\,.$$

(b) Let $U \subset X$ be a convex subset. A **function** *$\mathcal{J} : U \longrightarrow \mathbb{R}$ is called* **uniformly convex** *if there exists $c > 0$ such that*

$$\lambda\,\mathcal{J}(\psi_1) + (1-\lambda)\,\mathcal{J}(\psi_2) - \mathcal{J}\big(\lambda\psi_1 + (1-\lambda)\psi_2\big) \;\geq\; c\lambda(1-\lambda)\,\|\psi_1 - \psi_2\|^2 \tag{3.6}$$

for all $\psi_1, \psi_2 \in U$ and $\lambda \in [0,1]$.

(c) The function $\mathcal{J}$ is called **convex** *if (3.6) holds for $c = 0$.*

(d) The function is called **strictly convex** *if (3.6) holds strictly for $c = 0$ i.e.,*

$$\mathcal{J}\big(\lambda\psi_1 + (1-\lambda)\psi_2\big) \;<\; \lambda\,\mathcal{J}(\psi_1) \;+\; (1-\lambda)\,\mathcal{J}(\psi_2)$$

for all $\psi_1 \neq \psi_2$ and $\lambda \in (0,1)$.

Then we have

Theorem 3.5. *Every continuous and convex functional $\mathcal{J}$ on a Hilbert space X is weakly sequentially lower semi-continuous.*

Proof: Let the sequence $\{\psi_k\}_{k=1}^\infty$ converge weakly to some $\psi^o \in X$ and assume, on the contrary, that

$$\liminf_{k\to\infty} \mathcal{J}(\psi_k) \;<\; \mathcal{J}(\psi^o)\,.$$

Then there exists $\epsilon > 0$ and a subsequence $\bigl\{\psi_{k_j}\bigr\}_{j=1}^\infty$ of $\{\psi_k\}_{k=1}^\infty$ such that

$$\mathcal{J}(\psi^o) \;\geq\; \mathcal{J}\bigl(\psi_{k_j}\bigr) \;+\; \epsilon \quad \text{for all } j \in \mathbb{N}\,.$$

This implies that $\psi_{k_j} \in A$ where

$$A \;=\; \bigl\{\psi \in X : \mathcal{J}(\psi) \leq \mathcal{J}(\psi^o) - \epsilon\bigr\}\,.$$

The continuity and convexity of $\mathcal{J}$ imply that A is closed and convex. Furthermore, $\psi^o \notin A$. The strict separation theorem (Theorem A.46 of the Appendix) yields the existence of $c \in \mathbb{R}$ and $\phi \in X$ with

$$\mathrm{Re}\,(\psi^o, \phi)_X \;<\; c \;\leq\; \mathrm{Re}\,(\psi, \phi)_X \quad \text{for all } \psi \in A\,.$$

Taking $\psi = \psi_{k_j}$ yields

$$\mathrm{Re}\,(\psi^o, \phi)_X \;<\; c \;\leq\; \mathrm{Re}\,(\psi_{k_j}, \phi)_X \quad \text{for all } j \in \mathbb{N}\,,$$

and letting j tend to infinity yields $\mathrm{Re}\,(\psi^o, \phi)_X < c \leq \mathrm{Re}\,(\psi^o, \phi)_X$, a contradiction. □

Remark: We caution the reader that the hypothesis of convexity is crucial to this result. In an infinite dimensional Hilbert space functionals which are continuous (with respect to the norm) are *not* necessarily weakly sequentially lower semi-continuous as can be seen by simply considering $X = \ell^2$ and $\mathcal{J}(\psi) = -\|\psi\|_{\ell^2}$. Indeed, the sequence $\bigl\{\psi^{(j)}\bigr\}$ of unit sequences (which consist of zeros except at the $j-$component where it is one) converges weakly to zero but $\mathcal{J}\bigl(\psi^{(j)}\bigr) = -1$.

While these results insure that optimal solutions exist, there is no guarantee in general that such an optimal solution is unique. Indeed, several of the cost functionals we consider involve a quadratic form and simple examples show that such problems may well have multiple solutions. However, uniqueness holds if U is convex and $\mathcal{J}$ is strictly convex.

Theorem 3.6. *Let $U \subset X$ be convex and $\mathcal{J} : U \longrightarrow \mathbb{R}$ be strictly convex. Then there exists at most one minimum of $\mathcal{J}$ on U.*

Proof: Assume, on the contrary, that $\psi^o \in U$ and $\psi^{oo} \in U$ are two minima of $\mathcal{J}$ on U. From the strict convexity we conclude that

$$\mathcal{J}\Bigl(\frac{1}{2}\psi^o + \frac{1}{2}\psi^{oo}\Bigr) \;<\; \frac{1}{2}\,\mathcal{J}(\psi^o) \;+\; \frac{1}{2}\,\mathcal{J}(\psi^{oo}) \;=\; \mathcal{J}^*\,,$$

where $\mathcal{J}^* = \inf_{\psi \in U} \mathcal{J}(\psi)$. Therefore, $(\psi^o + \psi^{oo})/2$ yields a smaller value than the optimal solutions ψ^o and ψ^{oo} which is impossible. □

Analogously to Theorem 3.5, closedness with respect to the norm-topology plus a convexity property yields weak sequential closedness.

Theorem 3.7. *Let U be a non-empty, closed, convex subset of a normed space X. Then the set U is weakly sequentially closed. If U is also bounded and X is a reflexive Banach space (e.g. a Hilbert space) then U is weakly sequentially compact.*

Proof: Suppose that the sequence $\{x_k\}_{k=1}^{\infty} \subset U$ converges weakly to some $x_o \in X$. Hence $x^*(x_k) \to x^*(x_o)$ as k tends to infinity for any $x^* \in X^*$. We wish to show that $x_o \in U$. Suppose, on the contrary, that $x_o \notin U$. Then, by Theorem A.46 there exists an $\hat{x}^* \in X^*$ and a constant $c \in \mathbb{R}$ such that

$$\operatorname{Re}\big[\hat{x}^*(x_o)\big] \;<\; c \;\leq\; \operatorname{Re}\big[\hat{x}^*(x)\big] \quad \text{for all } x \in U\,.$$

Substituting $x = x_k \in U$ into this chain of inequalities yields

$$\operatorname{Re}\big[\hat{x}^*(x_o)\big] \;<\; c \;\leq\; \operatorname{Re}\big[\hat{x}^*(x_k)\big] \quad \text{for all } k \in \mathbb{N}\,.$$

For $k \to \infty$ this leads to the contradiction $\operatorname{Re}\big[\hat{x}^*(x_o)\big] < c \leq \operatorname{Re}\big[\hat{x}^*(x_o)\big]$. Now let, in addition, U be bounded and X be a reflexive Banach space. Then any sequence $\{x_k\}_{k=1}^{\infty} \subset U$ is bounded and, therefore, contains a weakly convergent subsequence by Theorem A.56. □

3.2.2 The Modeling of Constraints

In this subsection we study more concrete realizations of the set U of constraints. First we remark that, by Theorem 3.7, any convex, closed, and bounded set in a Hilbert space is weakly sequentially compact. This assumption is satisfied for many "simple" constraints as, e.g., power constraints or sign conditions. More complicated quantities as, e.g., the super-gain ratio (see below) can be rewritten in terms of inequalities of the form $g(\psi) \leq 0$. Combining these with the simple constraints leads to sets U of the form

$$U \;=\; \big\{\psi \in U_0 : g_i(\psi) \leq 0,\ i = 1, \dots, p\big\} \tag{3.7}$$

where $U_0 \subset X$ describes the simple constraints.

Theorem 3.8. *Let X be a normed space, $U_0 \subset X$ weakly sequentially closed, and $g_i : X \longrightarrow \mathbb{R}$, $i = 1, \dots, p$, weakly sequentially lower semi-continuous functions. Then the set U, defined by (3.7), is weakly sequentially closed. If, in addition, $U_0 \subset X$ weakly sequentially compact, and X is reflexive then also $U \subset X$ is weakly sequentially compact.*

In particular, if X is a Hilbert space sets U of the form (3.7) are weakly sequentially compact by Theorem 3.7 and Theorem 3.5 if $U_0 \subset X$ is bounded, closed and convex, and all g_i are continuous and convex.

Proof: It suffices to show that sets of the form

$$\hat{U} \;:=\; \{\psi \in X : g_i(\psi) \le 0\}$$

(for any $i = 1,\dots,p$) are weakly sequentially closed. Let $\{\psi_k\}_{k=1}^{\infty} \subset \hat{U}$ converge weakly to some $\psi \in X$. From the lower semi-continuity of g_i we conclude that $0 \ge \liminf_{j\to\infty} g_i\big(\psi_{k_j}\big) \ge g_i(\psi)$, i.e. $\psi \in \hat{U}$. □

We will describe two types of sets U_0 of elementary constraints. First, it is evident that practical considerations will prevent the creation of surface currents with very large power. Thus, if the space X is $L^2(\Gamma)$, we denote the currents again by ψ and consider only surface currents satisfying a condition of the form $\int_\Gamma |\psi(\boldsymbol{x})|^2\, dS \le M^2$ where M is a fixed constant. In this case,

$$U_0 \;=\; \{\psi \in X : \|\psi\|_X \;\le\; M\}\,. \tag{3.8}$$

This is a simple power constraint and defines a ball in X which is certainly a closed, bounded, and convex set. More generally, we could consider constraints described by uniformly convex inequalities.

Lemma 3.9. *Let X be a Hilbert space and $g : X \longrightarrow \mathbb{R}$ be uniformly convex and continuous. Then the set*

$$U_0 \;=\; \{\psi \in X : g(\psi) \;\le\; 0\} \tag{3.9}$$

is bounded, closed and convex and therefore weakly sequentially compact.

Proof: The closedness and convexity of U_0 follows immediately from, respectively, the continuity and convexity of g. It remains to show that U_0 is a bounded set. To this end, fix $\psi_0 \in U_0$ and let $\delta > 0$ such that $|g(\psi) - g(\psi_0)| \le 1$ for all ψ satisfying $\|\psi - \psi_0\| \le \delta$. Define $M := |g(\psi_0)| - 1$ and observe that $g(\psi) \ge M$ for all $\psi \in X$ with $\|\psi - \psi_0\| \le \delta$. Let $\psi \in U_0$, $\psi \ne \psi_0$, be arbitrary and set $\lambda = \delta / \|\psi - \psi_0\|$. We consider two cases. If $\lambda \ge \frac{1}{2}$ then $\|\psi - \psi_0\| \le 2\delta$. If $\lambda \le \frac{1}{2}$ then we use $1 - \lambda \ge 1/2$ and the uniform convexity and have

$$\begin{aligned}
0 &\ge \lambda g(\psi) \;+\; (1-\lambda) g(\psi_0) \\
&\ge g\big(\lambda\psi + (1-\lambda)\psi_0\big) \;+\; c\lambda(1-\lambda)\,\|\psi - \psi_0\|^2 \\
&\ge M \;+\; c\lambda(1-\lambda)\,\|\psi - \psi_0\|^2 \;=\; M \;+\; c\delta(1-\lambda)\,\|\psi - \psi_0\| \\
&\ge M \;+\; \frac{c\delta}{2}\,\|\psi - \psi_0\|
\end{aligned}$$

since

$$\|\lambda\psi + (1-\lambda)\psi_0 - \psi_0\| \;=\; \lambda\,\|\psi - \psi_0\| \;=\; \delta\,.$$

This proves that

$$\|\psi - \psi_0\| \;\le\; \max\{2\delta, -2M/(c\delta)\} \quad \text{for all } \psi \in U_0$$

and ends the proof. □

This lemma not only includes the case of (3.8) which we can rewrite as $g(\psi) = \|\psi\|_X^2 - M^2$, but also others. For example, we may wish to bound both the energy in ψ and also its oscillations. One simple approach to this requirement is to choose some **Sobolev space** $H^s(\Gamma)$ (see Appendix, Section A.3.2) for X and specify a bound of the type

$$\|\psi\|_{H^s(\Gamma)} \;\le\; M\,.$$

Of a completely different nature are **pointwise constraints** on ψ in the case that X is some function space. We formulate this class of constraints only for the case where the underlying space is $L^2(\Gamma, \mathbb{C}^q)$ for some compact set $\Gamma \subset \mathbb{R}^d$ and $q \in \mathbb{N}$ although it is possible to extend these considerations also to Sobolev spaces. $L^2(\Gamma, \mathbb{C}^q)$ is the space of vector fields $\psi : \Gamma \longrightarrow \mathbb{C}^q$ such that every component ψ_j of ψ is in $L^2(\Gamma)$.

Lemma 3.10. *Let $\Gamma \subset \mathbb{R}^d$ and for each $\boldsymbol{x} \in \Gamma$ let $V(\boldsymbol{x}) \subset \mathbb{C}^q$ be a closed and convex set such that $\bigcup_{\boldsymbol{x}\in\Gamma} V(\boldsymbol{x})$ is bounded. Let $U_0 \subset L^2(\Gamma, \mathbb{C}^q)$ be defined by*

$$U_0 \;:=\; \big\{\psi \in L^2(\Gamma, \mathbb{C}^q) : \psi(\boldsymbol{x}) \in V(\boldsymbol{x}) \text{ a.e. on } \Gamma\big\}\,. \tag{3.10}$$

Then $U_0 \subset L^2(\Gamma, \mathbb{C}^q)$ is convex, closed, and bounded in $L^2(\Gamma, \mathbb{C}^q)$ and thus weakly sequentially compact.

Proof: The convexity of the set U_0 follows immediately from the convexity of the individual sets $V(\boldsymbol{x})$. To see that U_0 is closed, suppose that $\{\psi_k\}_{k=1}^\infty \subset U_0$ is a sequence with $\psi_k \to \psi$ in $L^2(\Gamma, \mathbb{C}^q)$. Then there exists a subsequence, which we will again call $\{\psi_k\}_{k=1}^\infty$ which converges to the same function ψ pointwise almost everywhere in Γ. Let $\mathcal{N}_0$ be the set of measure zero where this convergence fails, and let $\mathcal{N}_k$ be the set where $\psi_k(\boldsymbol{x}) \notin V(\boldsymbol{x})$. Then $\mathcal{N} := \bigcup\{\mathcal{N}_k : k = 0, 1, \ldots\}$ is a set of measure zero, and $\psi_k(\boldsymbol{x}) \in V(\boldsymbol{x})$ for all $k \ge 0$ and $\boldsymbol{x} \notin \mathcal{N}$. Since the set $V(\boldsymbol{x})$ is closed, $\psi(\boldsymbol{x}) \in V(\boldsymbol{x})$ for all $\boldsymbol{x} \notin \mathcal{N}$. This proves that U_0 is closed. The boundedness of the set U_0 follows from the boundedness of $\bigcup_{\boldsymbol{x}\in\Gamma} V(\boldsymbol{x})$. □

We remark that in order to guarantee that $U_0 \neq \emptyset$ it is necessary to impose some continuity properties on the set-valued mapping $\boldsymbol{x} \mapsto V(\boldsymbol{x})$ i.e., of a map from the set Γ to *non-empty subsets* of $\mathbb{C}^q$.

Since set-valued mappings may not be familiar to the reader, we pause to introduce a definition and a background remark. The introduction of such objects can be quite useful in optimization theory where they can be used effectively to study continuous dependence results and sensitivity results for the optimal solutions of parametrized problems. They are also crucial in the study of the existence of optimal solutions. From the point of view of applications, they have been used with great effectiveness in the study of economic models and in the area of non-linear programming. Interested readers may consult, for example, [22] or [98].

In the single-valued case, there is a connection between the graph of the function and its continuity properties. In dealing with the more general notion of set-valued mappings we may well need some notions of "smooth dependence" of the sets on the independent variables and it is most convenient to make these definitions in terms of the *graph* of the function. For these reasons we find it convenient to introduce a definition.

Definition 3.11. *Let Γ be a subset of $\mathbb{R}^d$ and let $\mathcal{P}(\mathbb{C}^q)$ be the set of all subsets of $\mathbb{C}^q$. Let $Q : \Gamma \longrightarrow \mathcal{P}(\mathbb{C}^q) \setminus \{\emptyset\}$. Then the* **graph** *of Q, denoted by* Gr(Q) *is the set*

$$\mathrm{Gr}(Q) \;:=\; \{(\boldsymbol{x}, \boldsymbol{c}) \in \Gamma \times \mathbb{C}^q : \boldsymbol{c} \in Q(\boldsymbol{x})\}\,. \tag{3.11}$$

It is possible to put the structure of a metric space on the set $\mathcal{P}(\mathbb{C}^q)$, called the Hausdorff metric, but we will not find it necessary to do this. It is enough for many purposes in optimization theory to require that this set Gr(Q) is, say, a closed subset of $\mathbb{R}^d \times \mathbb{C}^q$ or that it is a Lebesgue measurable subset of this product space. Indeed, this latter requirement is the one used in the next result. Requirements of this type are necessary to guarantee that "selections" i.e., single valued functions $\varphi : \Gamma \longrightarrow \mathbb{C}^q$ with the property that $\varphi(\boldsymbol{x}) \in Q(\boldsymbol{x})$, have nice properties as, for example, continuity or integrability. Theorems of this type are called selection theorems (see e.g., [22]).

With these ideas in hand we return to the set (3.10) and give some examples of the occurrence of this type of very general constraint.

Example 3.12. Let $\Gamma \subset \mathbb{R}^d$.

(a) Let α^+, α^-, β^+, $\beta^- : \Gamma \longrightarrow \mathbb{R}^q$ be functions with values in $\mathbb{R}^q$. The subset $V(\boldsymbol{x})$ of $\mathbb{C}^q$ defined by

$$V(\boldsymbol{x}) \;:=\; \left\{ \boldsymbol{z} \in \mathbb{C}^q : \begin{matrix} \alpha_j^-(\boldsymbol{x}) \le \mathrm{Re}\, z_j \le \alpha_j^+(\boldsymbol{x}), \\ \beta_j^-(\boldsymbol{x}) \le \mathrm{Im}\, z_j \le \beta_j^+(\boldsymbol{x}), \end{matrix} j = 1, \ldots, q \right\} \tag{3.12a}$$

is a hypercube in $\mathbb{C}^q$, i.e. an q-fold product of rectangles in $\mathbb{C}$.

(b) Let $\alpha : \Gamma \longrightarrow \mathbb{R}^q$ and

$$V(\boldsymbol{x}) \;:=\; \{\boldsymbol{z} \in \mathbb{C}^q : |z_j| \le \alpha_j(\boldsymbol{x}),\; j = 1, \ldots, q\}\,. \tag{3.12b}$$

Then $V(\boldsymbol{x})$ is a product of disks in $\mathbb{C}$.

(c) Let $\delta : \Gamma \longrightarrow \mathbb{R}$ and $a_j > 0$, $j = 1, \ldots, q$, and

$$V(\boldsymbol{x}) := \left\{ \boldsymbol{z} \in \mathbb{C}^q : \sum_{j=1}^{q} a_j \, |z_j| \leq \delta(\boldsymbol{x}) \right\}. \tag{3.12c}$$

From this example we see that the abstract constraint given by (3.10) includes standard inequality constraints. The constraints in (b) and (c) can be summarized as $V(\boldsymbol{x}) = \{\boldsymbol{z} \in \mathbb{C}^q : \beta(\boldsymbol{z}, \boldsymbol{x}) \leq 0\}$ for some continuous function $\beta : \mathbb{C}^q \times \Gamma \longrightarrow \mathbb{R}$ which is strictly convex with respect to $\boldsymbol{z}$ for every $\boldsymbol{x} \in \Gamma$.

3.2.3 Extreme Points and Optimal Solutions

It is very useful to have some general idea of the *character* of the optimal points of a given functional. In the case where we are interested in *maximization* of a convex functional simple examples show that we have a useful characterization. For such an example, we need only turn to the maximization of the parabolic function $f(x, y) = x^2 + y^2$ on disks or rectangles. We expect the solution of the maximization problem (3.1) to lie on the boundary of the constraint set U and, even more, to be an extreme point of U. For example, in the case that the domain is a rectangle in $\mathbb{R}^2$, at least one of the corner points will be a point where the quadratic function f takes on its maximum. More generally, we can introduce a useful definition:

Definition 3.13. *Let X be a normed space and $U \subset X$. Then $\psi \in U$ is an* **extreme point** *of U if there is no $\varphi \in X$, $\varphi \neq 0$, such that $\psi \pm \varphi \in U$.*

In the case of a solid rectangle, it can easily be seen that the corner points are extreme points while, for the unit disk, all boundary points are extremal. Notice that both of these sets are closed, bounded, and convex. In the more general setting of an infinite dimensional Banach space, the existence of extremal points can be guaranteed under similar circumstances by a theorem of Krein and Milman (see [48], p. 362). This theorem is one of the fundamental principles in functional analysis.

In our present context that theorem asserts that any non-empty and "sufficiently compact" set has at least one extreme point. This fact will be quite useful to us when we discuss certain problems with constraint sets defined in terms of inequalities involving point values of given functions. It is also useful in treating the question of existence of optimal solutions as we will see below.

While the notion of extreme points is defined only in terms of the *linear* structure of the space X, it is interesting to note that the existence of an extreme point requires sufficiently strong topological assumptions on the set. This is the content of the Krein-Milman Theorem. Indeed, we have the following result which is a simplified version of the statement given in [48].

Theorem 3.14. *Let X be a reflexive Banach space*[3] *and suppose that a set $U \subset X$ is non-empty and weakly sequentially compact. Then there exist extreme points of U.*

We note that reflexivity of the space or a stronger compactness property is needed for this theorem. In the following example, the set U is convex, closed, bounded, and even sequentially closed but has no extreme points.

Example 3.15. Consider the unit ball $\mathcal{B} = B[0,1] \subset L^1(a,b)$. Certainly no element in the interior $B(0,1)$ of $\mathcal{B}$ can be an extreme point nor can any point of the boundary. Indeed, choose $f \in L^1(a,b)$ to be any function with $\|f\|_{L^1} = 1$. Let E be a set of positive measure on which $f \neq 0$ and let $\{E_1, E_2\}$ be disjoint sets of positive measure such that $E_1 \cup E_2 = E$. Let χ_{E_i}, $i = 1,2$, be the characteristic function of E_i. Define

$$f_i := \frac{f\,\chi_{E_i}}{\|f\,\chi_{E_i}\|}, \quad i = 1,2 .$$

Then, for each i, $f_i \neq f$ and $\|f_i\| = 1$ and so $f_i \in \mathcal{B}_1$, $i = 1,2$, and

$$\|f\,\chi_{E_1}\|\,f_1 \;+\; \|f\,\chi_{E_2}\|\,f_2 \;=\; f\,\chi_{E_1} \;+\; f\,\chi_{E_2} \;=\; f\,\chi_E \;=\; f\,, \tag{3.13}$$

since $E = E_1 \cup E_2$. Moreover,

$$\|f\,\chi_{E_1}\| + \|f\,\chi_{E_2}\| \;=\; \int_{E_1} |f(t)|\,dt + \int_{E_2} |f(t)|\,dt \;=\; \int_{E} |f(t)|\,dt \;=\; 1\,,$$

since $E_1 \cap E_2 = \emptyset$. So (3.13) is a non-trivial convex combination of $f_1, f_2 \in \mathcal{B}$. Hence f is *not* an extreme point. We conclude that $\mathcal{B} \subset L^1(a,b)$ has no extreme points although it is convex, closed, bounded, and weakly sequentially compact. We remark that this example does *not* work in any of the L^p spaces where $1 < p < \infty$ since the Banach spaces $L^p(a,b)$ for $1 < p < \infty$ are all reflexive in contrast to $L^1(a,b)$.

With these results from functional analysis, we can establish the promised theorem. Note that we consider the maximization of a convex functional rather than the minimization!

Theorem 3.16. *(a) Let X be a Hilbert space, U a bounded, closed, and convex subset, and let the functional $\mathcal{J} : U \longrightarrow \mathbb{R}$ be convex and weakly sequentially continuous. Then there exist optimal solutions of the optimization problem*

$$\text{Maximize} \quad \mathcal{J}(\psi) \quad \text{subject to} \quad \psi \in U\,, \tag{3.14}$$

and the optimal value is attained at an extreme point of U.

(b) If $\mathcal{J} : U \longrightarrow \mathbb{R}$ is strictly convex, $U \subset X$ convex, and $\psi^o \in U$ an optimal solution of problem (3.14), then ψ^o is necessarily an extreme point of U.

[3] Again, we note that we always assume that the space is also separable.

Remark: In part (a) we claim that there exists an extreme point which is an optimal solution. This does not rule out the possibility that other solutions exist which are *not* extreme points of U. Simple examples from linear programming show that, indeed, such solutions can exist. If strict convexity holds, however, then part (b) of this theorem shows that every optimal solution *must* be an extreme point of U.

Proof: (a) By Theorems 3.7 and 3.1 the set

$$\Phi \;:=\; \bigl\{\psi^o \in U : \psi^o \text{ is maximal for } \mathcal{J} \text{ on } U\bigr\} \tag{3.15}$$

is not empty. We show, first, that Φ is weakly sequentially compact. Since $\Phi \subset U$ and U itself is closed, bounded, and convex, and thus weakly sequentially compact (Theorem 3.7), it is sufficient to show that Φ is weakly sequentially closed. Indeed, consider a sequence $\{\psi_k\}_{k=1}^{\infty} \subset \Phi$ which converges weakly to some $\psi \in U$. Since $\mathcal{J}$ is weakly sequentially continuous, $\mathcal{J}(\psi_k)$ converges to $\mathcal{J}(\psi)$. From $\mathcal{J}(\psi_k) = \sup \mathcal{J}(U)$ we conclude that also $\mathcal{J}(\psi) = \sup \mathcal{J}(U)$, thus $\psi \in \Phi$. From this fact it follows that the set Φ is weakly sequentially compact. Theorem 3.14 asserts that there exist extreme points of Φ.

It remains to show that every extreme point of Φ is also an extreme point of U. To prove this, let $\psi^o \in \Phi$ be an extreme point of Φ. Let $\psi \in X$ with $\psi^o \pm \psi \in U$. Then

$$\mathcal{J}(\psi^o) \;=\; \mathcal{J}\Bigl(\frac{\psi^o+\psi}{2}+\frac{\psi^o-\psi}{2}\Bigr) \;\leq\; \frac{1}{2}\,\mathcal{J}(\psi^o+\psi)+\frac{1}{2}\,\mathcal{J}(\psi^o-\psi) \;\leq\; \mathcal{J}(\psi^o),$$

thus $\mathcal{J}(\psi^o+\psi) = \mathcal{J}(\psi^o-\psi) = \mathcal{J}(\psi^o)$. Therefore, $\psi^o \pm \psi \in \Phi$ which implies that ψ has to vanish since ψ^o is an extreme point of Φ. Hence ψ^o is also an extreme point of U and part (a) is proved.

We now turn to part (b). Let $\psi^o \in U$ be optimal and assume that ψ^o is *not* an extreme point of U. Then there exists a $\psi \in X$ with $\psi \neq 0$ for which $\psi^o \pm \psi \in U$. Since the functional $\mathcal{J}$ is strictly convex, we have the inequalities

$$\mathcal{J}(\psi^o) \;=\; \mathcal{J}\Bigl(\frac{\psi^o+\psi}{2}+\frac{\psi^o-\psi}{2}\Bigr) \;<\; \frac{1}{2}\,\mathcal{J}(\psi^o+\psi)+\frac{1}{2}\,\mathcal{J}(\psi^o-\psi) \;\leq\; \mathcal{J}(\psi^o),$$

which is a contradiction. This proves the assertion of part (b), and the theorem is established. □

For the case that U is the unit ball, the preceding result says that the optimal solution will be an extreme point of U. This is not necessarily the case for *minimizing* strictly convex functionals on bounded convex and closed sets. As a simple but relevant example of this latter statement consider the problem of feeding a single dipole so its far field matches a desired far field:

Example 3.17. Let $k > 0$, $\hat{\boldsymbol{y}} \in \mathbb{R}^3$, and $f \in L^2(S^2)$ be given and $\mathcal{K} : \mathbb{C} \longrightarrow C(S^2)$ be defined by

$$(\mathcal{K}a)(\hat{\boldsymbol{x}}) \;=\; a\,\mathrm{e}^{-ik\boldsymbol{y}\cdot\hat{\boldsymbol{x}}}\,, \quad \hat{\boldsymbol{x}} \in S^2 \subset \mathbb{R}^3\,,\; a \in \mathbb{C}\,. \tag{3.16}$$

The problem is to minimize the L^2–norm of $(\mathcal{K}a - f)$ with respect to $|a| \le 1$. In this case, the cost functional is given by

$$\begin{aligned}\mathcal{J}(a) = \|\mathcal{K}a - f\|^2_{L^2(S^2)} \;&=\; \int_{S^2} \left|a\,\mathrm{e}^{-ik\boldsymbol{y}\cdot\hat{\boldsymbol{x}}} - f(\hat{\boldsymbol{x}})\right|^2 dS(\hat{\boldsymbol{x}})\\ &= 4\pi\,|a|^2 \;-\; 8\pi\,\mathrm{Re}\,[a\,\overline{c}] \;+\; \|f\|^2_{L^2(S^2)}\\ &= 4\pi\,|a - c|^2 \;+\; \|f\|^2_{L^2(S^2)} \;-\; 4\pi\,|c|^2\,,\end{aligned}$$

where $c = \frac{1}{4\pi}\int_{S^2} \exp(ik\boldsymbol{y}\cdot\hat{\boldsymbol{x}})\,f(\hat{\boldsymbol{x}})\,dS(\hat{\boldsymbol{x}})$. From this we conclude that $a^o = c$ is the unique minimum of $\mathcal{J}$ on $U = \{z \in \mathbb{C} : |z| \le 1\}$ if $|c| < 1$ which is in the interior of U. If $|c| \ge 1$ then the unique solution is $a^o = c/|c|$ which lies on the boundary of U.

In order to illustrate the preceeding ideas, we consider, next, the particular types of constraint sets that we introduced in Subsection 3.2.2 and which we use in specific applications later in Chapter 7. In particular we look first at the case when the constraint set is given in terms of a finite number of inequalities involving real-valued functionals defined on X. In some sense, this is the classical case in optimization familiar from applications of non-linear programming. As a second example, we treat the case of pointwise constraints in the general form of so-called unilateral constraints, in which the conditions are given pointwise almost everywhere on the domain Γ. The nature of the extreme points is very different in these two cases as is the nature of the optimal solutions.

We begin by taking the set U to be of the form

$$U \;:=\; \bigl\{\psi \in X : g(\psi) \;\le\; 0\bigr\}\,, \tag{3.17}$$

where $g : X \dashrightarrow \mathbb{R}$ is some continuous and uniformly convex function. Then we know from Lemma 3.9 that U is closed, convex, and bounded. More concretely, we could take $g(\psi) = \|\psi\|^2_X - 1$ in which case U reduces to the unit ball in X. It is very easy to show that the extreme points of U are just the boundary points of U:

Lemma 3.18. *Let X be a Hilbert space, $g : X \longrightarrow \mathbb{R}$ continuous and strictly convex, and U given by (3.17). Then the set* ext U *of extreme points of U coincides with the set, ∂U, of boundary points, i.e.*

$$\mathrm{ext}\,U \;=\; \partial U \;=\; \bigl\{\psi \in X : g(\psi) = 0\bigr\}\,. \tag{3.18}$$

Proof: Let $\psi \in \mathrm{ext}\,U$ and assume on the contrary that $g(\psi) < 0$. From the continuity of ψ we conclude that $g(\psi \pm \varphi) \le 0$ for sufficiently small $\|\varphi\|$. This contradicts the assumption that ψ is an extreme point. Conversely, suppose $\psi \in \partial U$ and assume that there exists $\varphi \in X$ with $\varphi \neq 0$ and $\psi \pm \varphi \in U$ i.e.,

$g(\psi \pm \varphi) \leq 0$. Then, since g is strictly convex, $0 = g(\psi) = g\left(\frac{\psi+\varphi}{2} + \frac{\psi-\varphi}{2}\right) < \frac{1}{2}g(\psi+\varphi) + \frac{1}{2}g(\psi-\varphi) \leq 0$, which is a contradiction. This proves that ψ is an extreme point. □

We can conclude, in light of this result, that we can concentrate on finding optimal solutions which lie on the boundary of the constraint set i.e., for admissible functions on which the constraints are active. In particular, in the case that we take the simple but important example

$$g(\psi) \;:=\; \|\psi\|_X^2 \;-\; 1\,,$$

we can expect that we will obtain optimal solutions of unit norm in X.

Let us contrast this situation with the case that the constraint sets U involve pointwise constraints such as, e.g. sign conditions on ψ. We must first specify the underlying Hilbert space. For simplicity we take $X = L^2(\Gamma, \mathbb{C}^q)$ where $\Gamma \subset \mathbb{R}^d$, $d = 2$ or 3, is the C^2–boundary of an open and bounded set with connected exterior. We then define the set U by

$$U \;:=\; \left\{\boldsymbol{\psi} \in L^2(\Gamma, \mathbb{C}^q) : \boldsymbol{\psi}(\boldsymbol{x}) \in V(\boldsymbol{x}) \text{ a.e. on } \Gamma\right\} \tag{3.19}$$

where $\boldsymbol{x} \mapsto V(\boldsymbol{x}) \subset \mathbb{C}^q$ is some set-valued function defined for $\boldsymbol{x} \in \Gamma$.

In applications to antenna theory, we will study specific choices of this set-valued map V. Here we wish to characterize the extreme points of the set U, defined by a set-valued function, under only a mild regularity condition. Specifically, we can prove:

Lemma 3.19. *Let the set-valued function $\boldsymbol{x} \mapsto V(\boldsymbol{x}) \in \mathbb{C}^q$, defined on Γ, have a Lebesgue-measurable graph and define U by (3.19). Then*

$$\operatorname{ext} U \;=\; \left\{\boldsymbol{\psi} \in L^2(\Gamma, \mathbb{C}^q) : \boldsymbol{\psi}(\boldsymbol{x}) \in \operatorname{ext} V(\boldsymbol{x}) \textit{ a.e. on } \Gamma\right\}.$$

Proof: From the definition of an extreme point it is clear that the set on the right is contained in $\operatorname{ext} U$. Now let ψ be an extreme point of U and define the set-valued map

$$W(\boldsymbol{x}) \;:=\; \left\{\boldsymbol{z} \in D : \boldsymbol{\psi}(\boldsymbol{x}) \pm \boldsymbol{z} \in V(\boldsymbol{x})\right\} \quad \text{for } \boldsymbol{x} \in \Gamma\,,$$

where D is the poly-disk $\left\{\boldsymbol{z} \in \mathbb{C}^q : |z_j| \leq 1,\ j = 1, \ldots, q\right\}$. Then the graph $\operatorname{Gr}(W)$ of W is measurable since

$$\operatorname{Gr}(W) \;=\; (\Gamma \times D) \;\cap\; \operatorname{Gr}(V - \boldsymbol{\psi}) \;\cap\; \operatorname{Gr}(\psi - V)$$

and the graphs of $\pm(V - \boldsymbol{\psi})$ are measurable in $\Gamma \times \mathbb{C}^q$. The proof will be complete if we can show that $W(\boldsymbol{x}) = \{0\}$ almost everywhere on Γ.

Assume the contrary, that is, assume that there is a set of positive measure $I \subset \Gamma$ on which $W(\boldsymbol{x}) \neq \{0\}$. Using a measurable selection theorem (see [22]) we can choose a measurable function $\boldsymbol{\varphi}$, defined on Γ, with $\boldsymbol{\varphi}(\boldsymbol{x}) \in W(\boldsymbol{x})$ and $\boldsymbol{\varphi}(\boldsymbol{x}) \neq 0$ for all $\boldsymbol{x} \in I$. Since this selection is bounded, $\boldsymbol{\varphi} \in L^2(\Gamma, \mathbb{C}^q)$, $\boldsymbol{\varphi} \not\equiv 0$, and $\boldsymbol{\psi} \pm \boldsymbol{\varphi} \in U$ this contradicts the fact that $\boldsymbol{\psi} \in \operatorname{ext} U$, and the proof is complete. □

Example 3.20. Let the set $V(\boldsymbol{x})$ be given as

$$V(\boldsymbol{x}) := \left\{ \boldsymbol{z} \in \mathbb{C}^q : \begin{array}{l} \alpha_j^-(\boldsymbol{x}) \leq \operatorname{Re} z_j \leq \alpha_j^+(\boldsymbol{x}), \\ \beta_j^-(\boldsymbol{x}) \leq \operatorname{Im} z_j \leq \beta_j^+(\boldsymbol{x}), \end{array} \; j = 1, \dots, q \right\}, \tag{3.20a}$$

or

$$V(\boldsymbol{x}) := \left\{ \boldsymbol{z} \in \mathbb{C}^q : \beta(\boldsymbol{z}, \boldsymbol{x}) \leq 0 \right\} \tag{3.20b}$$

for some $\alpha^\pm, \beta^\pm : \Gamma \longrightarrow \mathbb{R}^q$ and some continuous function $\beta : \mathbb{C}^q \times \Gamma \longrightarrow \mathbb{R}$ which is assumed to be strictly convex with respect to $\boldsymbol{z}$ for every $\boldsymbol{x} \in \Gamma$. Then the extreme points are those which satisfy

$$z_j = \alpha_j^\sigma(\boldsymbol{x}) + i\,\beta_j^\rho(\boldsymbol{x}), \; j = 1, \dots, q, \; \sigma, \rho \in \{+, -\}, \quad \text{or} \quad \beta(\boldsymbol{z}, \boldsymbol{x}) = 0,$$

respectively, for almost all $\boldsymbol{x} \in \Gamma$.

3.2.4 The Lagrange Multiplier Rule

One of the important strategies in dealing with optimization problems is to develop necessary conditions for optimal solutions. As in the case of finite-dimensional constrained optimization problems, the use of the *Lagrange multiplier rule* is often very convenient in this regard. Its formulation involves the derivatives of the cost functional $\mathcal{J} : X \longrightarrow \mathbb{R}$ and the functionals which describe the constraint set, see Section A.7. We note that we consider here X as a Hilbert space over $\mathbb{R}$ - even if it is, e.g., a space of complex-valued functions. The **Fréchet derivative** of such a functional $\mathcal{J}$ at $\psi^o \in X$ is described by the **gradient** $\nabla \mathcal{J}(\psi^o) \in X$ through

$$\frac{1}{\|\varphi\|_X} \left[\mathcal{J}(\psi^o + \varphi) - \mathcal{J}(\psi^o) - \operatorname{Re}\left(\nabla \mathcal{J}(\psi^o), \varphi \right)_X \right] \longrightarrow 0$$

as $\|\varphi\|_X$ tends to zero.

First, we prove a familiar necessary condition for an optimal solution in the form of a **variational inequality**.

Lemma 3.21. *Let X be a Hilbert space, $U \subset X$ be convex and closed, and $\mathcal{J} : X \longrightarrow \mathbb{R}$. Let $\psi^o \in U$ be a minimum of $\mathcal{J}$ on U. Let $\mathcal{J}$ be continuously Fréchet differentiable in ψ^o with gradient $\nabla \mathcal{J}(\psi^o) \in X$. Then*

$$\operatorname{Re}\left(\nabla \mathcal{J}(\psi^o), \psi - \psi^o \right)_X \geq 0 \quad \textit{for all } \psi \in U. \tag{3.21}$$

Proof: Let $\psi \in U$. Then $\psi_\lambda := \psi^o + \lambda(\psi - \psi^o) \in U$ for every $\lambda \in (0, 1]$. Using the optimality of ψ^o we conclude that

$$\begin{aligned} 0 &\leq \frac{1}{\lambda} \left[\mathcal{J}(\psi_\lambda) - \mathcal{J}(\psi^o) \right] \\ &= \operatorname{Re}\left(\nabla \mathcal{J}(\psi^o), \psi - \psi^o \right)_X + \frac{1}{\lambda} \left[\mathcal{J}(\psi_\lambda) - \mathcal{J}(\psi^o) - \operatorname{Re}\left(\nabla \mathcal{J}(\psi^o), \psi_\lambda - \psi^o \right)_X \right]. \end{aligned}$$

Letting λ tend to zero yields the desired inequality (3.21) since, by the differentiability of $\mathcal{J}$,

$$\frac{1}{\lambda}\left[\mathcal{J}(\psi_\lambda) - \mathcal{J}(\psi^o) - \operatorname{Re}\left(\nabla\mathcal{J}(\psi^o), \psi_\lambda - \psi^o\right)_X\right] \longrightarrow 0,$$

as λ tends to zero. □

For the Lagrange multiplier rule it is assumed that the set U can be described by equations or inequalities. We restrict ourselves to the case of inequalities, i.e. $U \subset X$ is described as

$$U = \left\{\psi \in X : g_j(\psi) \leq 0, \quad j = 1, \ldots, m\right\},$$

where $g : X \longrightarrow \mathbb{R}^m$ is some vector valued function (see (3.7)). We call g **Fréchet differentiable** if every component $g_j : X \longrightarrow \mathbb{R}$ is differentiable.

Theorem 3.22. *Let X be a Hilbert space, $\mathcal{J} : X \longrightarrow \mathbb{R}$ and $g : X \longrightarrow \mathbb{R}^m$ be continuously Fréchet differentiable, and $\psi^o \in X$ be a solution of the optimization problem*

$$\text{Minimize} \quad \mathcal{J}(\psi) \quad \text{subject to} \quad \begin{cases} \psi \in X & \textit{and} \\ g_j(\psi) \leq 0, & j = 1, \ldots, m. \end{cases} \tag{3.22}$$

Let the following **constraint qualification** *be satisfied: There exists $\hat{\psi} \in X$ with*

$$g_j(\psi^o) + \operatorname{Re}\left(\nabla g_j(\psi^o), \hat{\psi}\right)_X < 0, \quad j = 1, \ldots, m. \tag{3.23}$$

Then there exist **Lagrange multipliers,** *i.e. real numbers $\rho_1, \ldots, \rho_m \geq 0$, such that*

$$\nabla\mathcal{J}(\psi^o) + \sum_{j=1}^{m} \rho_j \nabla g_j(\psi^o) = 0, \quad \textit{and} \tag{3.24a}$$

$$\rho_j \, g_j(\psi^o) = 0, \quad j = 1, \ldots, m. \tag{3.24b}$$

Remarks: This theorem assumes the existence of an optimal solution which must be assured by different methods e.g., by Theorems 3.1 or 3.3. The existence of Lagrange multipliers is only a *necessary* condition for ψ^o to be optimal; in general, this condition is not sufficient. Maximization problems, where one wishes to maximize a functional $\mathcal{J}$ instead of to minimize it, are obviously converted into minimization problems just by replacing $\mathcal{J}$ with $-\mathcal{J}$. The corresponding Lagrange multiplier rule differs only in that the plus sign in (3.24a) is now replaced by a minus sign.

The second condition (3.24b) implies that only multipliers ρ_j for active constraints have to be introduced. Here a constraint $g_j(\psi) \leq 0$ is called **active** for ψ^o if $g_j(\psi^o) = 0$. Indeed, if $g_j(\psi^o) < 0$ then (3.24b) implies that $\rho_j = 0$.

The form (3.23) of the constraint qualification can be written differently by using active and **inactive** constraints. Let $A(\psi^o) \subset \{1, \ldots, m\}$ be the set of the active constraints for ψ^o. Then (3.23) is equivalent to the existence of $\hat{\psi} \in X$ with

$$\operatorname{Re}\left(\nabla g_j(\psi^o), \hat{\psi}\right)_X \; < \; 0 \quad \text{for all } j \in A(\psi^o) .$$

Indeed, for $j \notin A(\psi^o)$ we have that $g_j(\psi^o) < 0$ and the inequality in (3.23) is always satisfied if we choose $\hat{\psi}$ small enough.

It will be useful to know that the constraint qualification (3.23), sometimes referred to as a generalized Slater condition, implies that there exist admissible functions on which the constraint is *inactive.*

Lemma 3.23. *Assume, that g is Fréchet differentiable in some $\psi^o \in X$, and that the constraint qualification (3.23) is satisfied for some $\hat{\psi} \in X$. There exists $\overline{\psi} \in X$ with*

$$g_j(\overline{\psi}) \; < \; 0, \quad j = 1, 2, \ldots, m . \tag{3.25}$$

Proof: For each integer j we consider the trivial identity

$$\begin{aligned} g_j\left(\psi^o + \varepsilon\hat{\psi}\right) = {} & \left[g_j(\psi^o + \varepsilon\hat{\psi}) - g_j(\psi^o) - \operatorname{Re}\left(\nabla g_j(\psi^o), \varepsilon\hat{\psi}\right)_X\right] \\ & + \varepsilon\left[g_j(\psi^o) + \operatorname{Re}\left(\nabla g_j(\psi^o), \hat{\psi}\right)_X\right] \; + \; (1-\varepsilon)\, g_j(\psi^o) . \end{aligned}$$

If we set

$$\begin{aligned} \eta_1(\varepsilon) &:= \max_{j=1,\ldots,m} \left|g_j(\psi^o + \varepsilon\hat{\psi}) - g_j(\psi^o) - \operatorname{Re}\left(\nabla g_j(\psi^o), \varepsilon\hat{\psi}\right)_X\right| \quad \text{and} \\ \eta_2 &:= \max_{j=1,\ldots,m} \left[g_j(\psi^o) \; + \; \operatorname{Re}\left(\nabla g_j(\psi^o), \hat{\psi}\right)_X\right] \; < \; 0 \end{aligned}$$

then we can make the estimate

$$\begin{aligned} g_j(\psi^o + \varepsilon\hat{\psi}) &\leq \eta_1(\varepsilon) \; + \; \varepsilon\,\eta_2 \; + \; (1-\varepsilon)\, g_j(\psi^o) \\ &= \varepsilon \underbrace{\left[\frac{\eta_1(\varepsilon)}{\varepsilon} \; + \; \eta_2\right]}_{<0} \; + \; (1-\varepsilon) \underbrace{g_j(\psi^o)}_{\leq 0} \; < \; 0 . \end{aligned}$$

The term in the bracket $[\cdots]$ is strictly less than zero for sufficiently small $\varepsilon > 0$ since $\lim_{\varepsilon\to 0}\left(\eta_1(\varepsilon)/\varepsilon\right) = 0$ by the definition of the derivative. This proves the lemma by taking $\overline{\psi} = \psi^o + \varepsilon\hat{\psi}$. □

Remark: As noted before, a constraint of the form $\|\psi\|_X \leq 1$ can be rewritten as an inequality $g_0(\psi) := \|\psi\|_X^2 - 1 \leq 0$. Since its gradient is $\nabla g_0(\psi^o) = 2\psi^o$ the constraint qualification (3.23) takes the form:

$$\begin{aligned} &\text{There exists } \hat{\psi} \in X \text{ with} \\ &g_j(\psi^o) \; + \; \operatorname{Re}\left(\nabla g_j(\psi^o), \hat{\psi}\right)_X \; < \; 0\,, j = 1, \ldots, m, \quad \text{and} \\ &\|\psi^o\|_X^2 \; + \; 2\operatorname{Re}\left(\psi^o, \hat{\psi}\right)_X \; < \; 1 . \end{aligned} \tag{3.26}$$

The results collected in this subsection form the basis of much of the numerical computations that we make in the remaining part of the book. But it is important to reiterate a fundamental point. The use of Lagrange multipliers, or any other necessary condition is predicted on the fact that we know that an optimal solution exists.

3.2.5 Methods of Finite Dimensional Approximation

This subsection is devoted to a general method for the numerical treatment of the optimization problem (3.1). Again let X be a Hilbert space e.g., $L^2(\Gamma)$, and $X_n \subset X$ a sequence of finite dimensional subspaces. Here, we make only the assumption that these subspaces are **ultimately dense** in X (see [73]), i.e. that $X_n \subset X_{n+1}$ for all n and $\bigcup_n X_n$ is dense in X.

There are many concrete ways to generate such an ultimately dense sequence of *finite dimensional* (and therefore closed) subspaces. Experience shows that particular problems suggest a particular choice of this sequence of subspaces, and that choice may well effect the convergence of the solutions of the finite dimensional problems to a solution of the full infinite dimensional one. For example, in the case $X = L^2(\Gamma)$ where Γ is the smooth boundary curve of a plane region, we can think of X_n as being spaces of trigonometric functions (with respect to the parameterization of the curve Γ), or spaces of continuous, piecewise polynomials (boundary elements). It is, however, also possible to use a family of linearly independent solutions of the underlying differential equation without necessarily satisfying any additional (e.g., boundary) condition and which is *complete* or *fundamental* in the space X. We refer to Subsection 5.5.2 for examples. These methods can be used to solve both, the boundary value problem and the optimization problem. We refer again to Subsection 5.5.2 and Chapter 7 for more details.

Now we pose a finite dimensional version of the optimization problem (3.1):

$$\text{Minimize} \quad \mathcal{J}(\psi) \quad \text{subject to} \quad \psi \in X_n \cap U\,. \tag{3.27}$$

Before we prove the main theorem we need the following approximation result with respect to the sets $X_n \cap U$:

Lemma 3.24. *Let $X_n \subset X$ be a sequence of subspaces such that $\bigcup_n X_n$ is dense in X and, furthermore, let $U \subset X$ be a convex set with nonempty interior $\overset{o}{U}$. Then, for every $\hat{\psi} \in U$ there exists a sequence $\psi_n \in X_n \cap U$ with $\psi_n \to \hat{\psi}$ in X.*

Proof: Fix some $\psi_1 \in \overset{o}{U}$ and define $\psi_\lambda := \hat{\psi} + \lambda(\psi_1 - \hat{\psi})$ for $\lambda \in [0,1]$. First we show that $\psi_\lambda \in \overset{o}{U}$ for all $\lambda \in (0,1]$. Indeed, let $\epsilon > 0$ be such that $B(\psi_1, \epsilon) \subset U$ where again $B(\psi_1, \epsilon)$ denotes the open ball with center ψ_1 and radius ϵ. Then $B(\psi_\lambda, \epsilon\lambda) \subset U$ since if $\psi \in B(\psi_\lambda, \epsilon\lambda)$ then we observe from

$$\left\|\hat{\psi} + \frac{1}{\lambda}(\psi - \hat{\psi}) - \psi_1\right\| = \frac{1}{\lambda}\, \|\psi_\lambda - \psi\|$$

that $\hat{\psi} + \frac{1}{\lambda}(\psi - \hat{\psi}) \in B(\psi_1, \epsilon) \subset U$. Therefore,

$$\psi = \lambda\left[\hat{\psi} + \frac{1}{\lambda}(\psi - \hat{\psi})\right] + (1-\lambda)\,\hat{\psi} \in U\,.$$

Thus we have shown that $\hat{\psi} + \frac{1}{n}(\psi_1 - \hat{\psi}) \in \overset{o}{U}$ for every $n \in \mathbb{N}$. By the hypotheses of the lemma, for every $n \in \mathbb{N}$ there exists $m_n \in \mathbb{N}$ and $\psi_{m_n} \in X_{m_n}$ with

$$\left\|\hat{\psi} + \frac{1}{n}(\psi_1 - \hat{\psi}) - \psi_{m_n}\right\| \leq \frac{1}{n} \quad \text{and} \quad \psi_{m_n} \in U\,.$$

We can also assume that $\{m_n\}$ is a strictly increasing sequence. Therefore, $\psi_{m_n} \in X_{m_n} \cap U$ and

$$\left\|\hat{\psi} - \psi_{m_n}\right\| \leq \left\|\hat{\psi} + \frac{1}{n}(\psi_1 - \hat{\psi}) - \psi_{m_n}\right\| + \frac{1}{n}\left\|\psi_1 - \hat{\psi}\right\| \leq \frac{1}{n}(1 + \|\psi_1 - \hat{\psi}\|)$$

which tends to zero as n tends to infinity. Setting $\psi_p = \psi_{m_n} \in X_{m_n} \subset X_p$ for $m_n \leq p < m_{n+1}$ yields that $\psi_p \to \hat{\psi}$ as $p \to \infty$. This ends the proof. □

We note that the assumption that U has interior points is crucial for the construction. Sets defined by pointwise constraints as e.g., in (3.10), may well have empty interiors in L^2-spaces. This will require a different approximation scheme which we will discuss below. But first we continue with the main approximation result:

Theorem 3.25. *Assume that $\mathcal{J} : X \longrightarrow \mathbb{R}$ is continuous and weakly lower sequentially semi-continuous, the set U is closed, convex, and bounded with nonempty interior. Assume, furthermore, that $X_n \subset X$, $n = 1, 2, \ldots$, is a sequence of finite dimensional subspaces such that $X_n \subset X_{n+1}$ for all n and $\bigcup_n X_n$ is dense in X. Then, there exists $n_0 \in \mathbb{N}$ such that $X_n \cap U \neq \emptyset$ for all $n \geq n_0$, and the optimization problems (3.27) have optimal solutions for $n \geq n_0$. Furthermore, any sequence $\left\{\psi_n^o\right\}_{n=1}^{\infty} \subset X_n \cap U$ of optimal solutions has weak accumulation points and every such weak accumulation point is optimal for the minimization of $\mathcal{J}$ on U. Finally, the optimal values $\mathcal{J}_n^o := \min\{\mathcal{J}(\psi) : \psi \in X_n \cap U\}$ converge to the optimal value $\mathcal{J}^o := \min\{\mathcal{J}(\psi) : \psi \in U\}$.*

Proof: From the previous lemma, it follows that $X_n \cap U \neq \emptyset$ for sufficiently large n. Moreover, the existence of optimal solutions follows directly from the fact that $X_n \cap U$ is again weakly sequentially compact (and even compact since X_n is finite dimensional). Since $\{\psi_n^o\}_{n=1}^\infty \subset U$, and U is weakly sequentially compact, there exist weak accumulation points of $\{\psi_n^o\}_{n=1}^\infty$. Let $\psi^o \in U$ be such a weak accumulation point and $\psi_{n_j}^o \rightharpoonup \psi^o$ weakly in X as $j \to \infty$. Furthermore, let $\hat{\psi} \in U$ be any optimal solution of (3.2) and $\hat{\psi}_n \in (X_n \cap U)$

with $\hat{\psi}_n \to \hat{\psi}$ as $n \to \infty$. The existence of such a sequence is assured by the previous lemma. Then we consider the following chain of inequalities:

$$\begin{aligned} \mathcal{J}^o \leq \mathcal{J}(\psi^o) \ &\leq \ \liminf_{j\to\infty} \mathcal{J}\big(\psi^o_{n_j}\big) \ = \ \liminf_{j\to\infty} \mathcal{J}^o_{n_j} \\ &\leq \limsup_{j\to\infty} \mathcal{J}^o_{n_j} \ \leq \ \limsup_{j\to\infty} \mathcal{J}\big(\hat{\psi}_{n_j}\big) \ = \ \mathcal{J}(\hat{\psi}) \ = \ \mathcal{J}^o . \end{aligned}$$

This shows that equality holds everywhere and, in particular, that ψ^o is minimal for $\mathcal{J}$ on U and $\mathcal{J}^o_{n_j}$ converges to $\mathcal{J}^o$ as j tends to infinity. This implies convergence of the sequence $\{\mathcal{J}^o_n\}_{n=1}^\infty$ itself since it is monotonically nonincreasing. This ends the proof. □

Remark: Concerning the hypothesis on the functional $\mathcal{J}$, we remark that the requirement that $\mathcal{J}$ be continuous is *weaker* than the alternative hypothesis that $\mathcal{J}$ be weakly continuous. While the proof here, *mutatis mutandis*, is valid in the case that $\mathcal{J}$ is weakly continuous, we use the weaker hypothesis.

The approximation of the optimization problem formulated in (3.27) does not take into account the approximation of the functional $\mathcal{J}$ itself or of the constraint set U. In practice, however, both the cost functional and the constraint set must be approximated as well. Thus one must consider the optimization problem

$$\text{Minimize} \quad \mathcal{J}_n(\psi) \quad \text{subject to } \psi \in U_n \tag{3.28}$$

where $U_n \subset X_n$ is some approximation of $U \subset X$ and $\mathcal{J}_n : X_n \longrightarrow \mathbb{R}$ is some approximation of the original functional $\mathcal{J} : X \longrightarrow \mathbb{R}$. The assertion of Theorem 3.25 remains true (with the same proof) if the following set of conditions is satisfied:

(i) If $\psi_n \in U_n$, $\psi_n \rightharpoonup \psi$ weakly, then $\psi \in U$ and $\liminf_{n\to\infty} \mathcal{J}_n(\psi_n) \geq \mathcal{J}(\psi)$;
(ii) If $\psi_n \in U_n$, $\psi_n \to \psi$ strongly, then $\lim_{n\to\infty} \mathcal{J}_n(\psi_n) = \mathcal{J}(\psi)$;
(iii) Every U_n is closed and convex and $\bigcup_n U_n$ is bounded;
(iv) For every $\psi \in U$ there exist $\psi_n \in U_n$ with $\psi_n \to \psi$.

Before we investigate a different approximation scheme which does not require that the constraint set has interior points, we recall the notion of an **projection operator** on a convex set. Let $U \subset X$ be a closed and convex subset of a Hilbert space X. By Theorem A.13, for every $\psi \in X$ there exists a unique element $P(\psi) \in U$ with $\|P(\psi) - \psi\| \leq \|\varphi - \psi\|$ for all $\varphi \in U$. Also, $P(\psi) \in U$ is uniquely determined by the variational inequality

$$\operatorname{Re}\big(P(\psi) - \psi\,,\, P(\psi) - \varphi\big)_X \ \leq \ 0 \quad \text{for all } \varphi \in U\,. \tag{3.29}$$

This construction defines, in general, a *nonlinear*, continuous (even Lipschitzian) operator $P : X \longrightarrow U \subset X$. Indeed, substitute $\varphi_1 = P(\psi_2) \in U$ and $\varphi_2 = P(\psi_1) \in U$ in the inequalities

$$\operatorname{Re}\big(P(\psi_1)-\psi_1\,,\,P(\psi_1)-\varphi_1\big)\ \le\ 0 \quad\text{and}\quad \operatorname{Re}\big(\psi_2-P(\psi_2)\,,\,\varphi_2-P(\psi_2)\big)\ \le\ 0$$

and add the results. This yields $\operatorname{Re}\big(P(\psi_1)-\psi_1+\psi_2-P(\psi_2)\,,\,P(\psi_1)-P(\psi_2)\big) \le 0$ and thus, by the Cauchy-Schwarz inequality,

$$\begin{aligned}\|P(\psi_1)-P(\psi_2)\|_X^2 &\le \operatorname{Re}\,\big(\psi_1-\psi_2\,,\,P(\psi_1)-P(\psi_2)\big)\\ &\le \|P(\psi_1)-P(\psi_2)\|_X\ \|\psi_1-\psi_2\|_X\ .\end{aligned}$$

From this inequality we may conclude that $\|P(\psi_1)-P(\psi_2)\|_X \le \|\psi_1-\psi_2\|_X$ and, in particular, that P is continuous. Also, we note that $P(\varphi)=\varphi$ for all $\varphi\in U$.

We consider two examples to illustrate this idea of projection onto a convex set.

Examples 3.26.

(a) First, define U to be the ball with center 0 and radius $R>0$. Then $P: X \longrightarrow U \subset X$ is given by $P(\psi)=\min\{1, R/\|\psi\|\}\,\psi$ for $\psi\in X$. This is easily seen by checking that the variational inequality is satisfied for this choice of P. To do this, let $\|\psi\|\ge R$ and $\varphi\in U$ and set $\hat\psi = R\psi/\|\psi\|$. Then

$$\begin{aligned}\operatorname{Re}\,\big(\psi-\hat\psi\,,\,\varphi-\hat\psi\big) &= \big(1-R/\|\psi\|\big)\operatorname{Re}\,\big(\psi\,,\,\varphi-\hat\psi\big)\\ &= \big(1-R/\|\psi\|\big)\,\big[\operatorname{Re}\,(\psi,\varphi)-R\,\|\psi\|\big]\ \le\ 0\end{aligned}$$

by the Cauchy-Schwarz inequality and $\|\varphi\|\le R$. Therefore, $\hat\psi = P(\psi)$.

(b) As a second example we consider sets with empty interior of the form (3.10), i.e.

$$U\ :=\ \big\{\boldsymbol{\psi}\in L^2(\varGamma,\mathbb{C}^q) : \boldsymbol{\psi}(\boldsymbol{x})\in V \text{ a.e. on } \varGamma\big\}$$

where $V\subset\mathbb{C}^q$ is a closed and convex set. Then, for every $\boldsymbol{z}\in\mathbb{C}^q$ there exists the unique best approximation $Q(\boldsymbol{z})\in V$ of $\boldsymbol{z}$ in V. By integrating the variational inequality at the point $\boldsymbol{z}=\boldsymbol{\psi}(\boldsymbol{x})$ with respect to $\boldsymbol{x}$, it is immediately seen that the orthogonal projection $P: L^2(\varGamma,\mathbb{C}^q)\longrightarrow U\subset L^2(\varGamma,\mathbb{C}^q)$ is given by $P(\boldsymbol{\psi})(\boldsymbol{x}) = Q\big(\boldsymbol{\psi}(\boldsymbol{x})\big)$, $\boldsymbol{x}\in\varGamma$, $\boldsymbol{\psi}\in L^2(\varGamma,\mathbb{C}^q)$.

The nonlinear projection mapping P allows us to rewrite the optimization problem (3.1) as an unconstrained problem in the form

$$\text{Minimize}\quad \mathcal{J}\big(P(\psi)\big)\quad \text{on } X\,.$$

Furthermore, if we assume that U is bounded and contained in the ball with center 0 and radius $R>0$ then it is easily seen that we can even add the additional constraint $\|\psi\|_X\le R$, and so study the optimization problem

$$\text{Minimize}\quad \mathcal{J}\big(P(\psi)\big)\quad \text{subject to } \psi\in X\,,\ \|\psi\|_X\le R\,. \tag{3.30}$$

Indeed, the proof of the following lemma is very simple and left to the reader:

Lemma 3.27. *If $\psi^o \in X$ with $\|\psi^o\|_X \leq R$ solves (3.30) then $P(\psi^o) \in U$ is optimal for the minimization of $\mathcal{J}$ on U. If, on the other hand, $\psi^o \in U$ is optimal for the minimization of $\mathcal{J}$ on U then ψ^o also solves (3.30). Furthermore, the optimal values of both optimization problems coincide.*

We note that, in general, the performance functional $\mathcal{J} \circ P$ of (3.30) is no longer weakly sequentially lower semi-continuous since P is not weakly continuous in general. Nevertheless, the form (3.30) now suggests the following approximation scheme:

$$\text{Minimize} \quad \mathcal{J}\big(P(\psi)\big) \quad \text{subject to } \psi \in X_n\,, \ \|\psi\|_X \leq R\,. \tag{3.31}$$

We can prove the following theorem in a manner analogous to the proof of Theorem 3.25.

Theorem 3.28. *Assume, as before, that $\mathcal{J} : X \longrightarrow \mathbb{R}$ is continuous and weakly sequentially lower semi-continuous and that the set U is closed, convex, and bounded with $\|\psi\|_X \leq R$ for all $\psi \in U$. Assume, furthermore, that $X_n \subset X$, $n = 1, 2, \ldots$, is a sequence of finite dimensional subspaces such that $X_n \subset X_{n+1}$ for all n and $\bigcup_n X_n$ is dense in X. Then there exist optimal solutions of (3.31) for all $n \in \mathbb{N}$. If $\psi_n^o \in X_n$ is optimal for (3.31) then $\big\{P(\psi_n^o)\big\}_{n=1}^{\infty} \subset U$ contains weak accumulation points and every such weak accumulation point is optimal for the minimization of $\mathcal{J}$ on U. Again, the optimal values $\mathcal{J}_n^o := \min\big\{\mathcal{J}\big(P(\psi)\big) : \psi \in X_n, \ \|\psi\|_X \leq R\big\}$ converge to the optimal value $\mathcal{J}^o := \min\{\mathcal{J}(\psi) : \psi \in U\}$.*

Proof: The existence of optimal solutions follows directly from the fact that the performance functional is continuous and the constraint set $\big\{\psi \in X_n : \|\psi\|_X \leq R\big\}$ is compact since it is a closed and bounded set in a finite dimensional space. Since $\big\{P(\psi_n^o)\big\}_{n=1}^{\infty} \subset U$ and U is weakly compact there exist weak accumulation points of the sequence $\big\{P(\psi_n^o)\big\}_{n=1}^{\infty}$. Let $\psi^o \in U$ be such a weak accumulation point and $P(\psi_{n_j}^o) \rightharpoonup \psi^o$ weakly in X. Furthermore, again let $\hat{\psi} \in U$ be any optimal solution of (3.2) and $\hat{\psi}_n \in X_n$ be a sequence with $\big\|\hat{\psi}_n\big\| \leq R$ and which converges to $\hat{\psi}$. Then we consider the following chain of inequalities:

$$\begin{aligned}\mathcal{J}^o \leq \mathcal{J}(\psi^o) \ &\leq \ \liminf_{j\to\infty} \mathcal{J}\big(P(\psi_{n_j}^o)\big) \ = \ \liminf_{j\to\infty} \mathcal{J}_{n_j}^o \\ &\leq \limsup_{j\to\infty} \mathcal{J}_{n_j}^o \ \leq \ \limsup_{j\to\infty} \mathcal{J}\big(P(\hat{\psi}_{n_j})\big) \ = \ \mathcal{J}\big(P(\hat{\psi})\big) \ = \ \mathcal{J}(\hat{\psi}) \ = \ \mathcal{J}^o\,.\end{aligned}$$

The remaining parts of the proof are completed just as in the proof of Theorem 3.25. □

As before, in this theorem we have not taken into account the approximation of either $\mathcal{J}$ or of the set U. If we make the same assumptions (i) – (iv) as for problem (3.28), then we may consider the more realistic approximate optimization problem

$$\text{Minimize} \quad \mathcal{J}_n(P_{U_n}(\psi)) \quad \text{subject to } \psi \in X_n,\ \|\psi\|_X \leq R\,, \tag{3.32}$$

where P_{U_n} denotes the projection on U_n and R is the radius of a ball containing the union $\bigcup_n U_n$. The assertion of Theorem 3.28 carries over to this case if we make the proper assumptions on U_n and $\mathcal{J}_n$.

3.3 Far Field Patterns and Far Field Operators

Up to this point, we have discussed constrained optimization problems in a very general way. Our ultimate goal is to *apply* the results of that discussion to problems of electromagnetic radiation. Our approach to such applications is to pose optimization problems in terms of functionals which relate the boundary data on a radiating structure to the far field pattern produced by that boundary data. In other words, the functionals to be optimized will be expressed in terms of the far field operator $\mathcal{K}$ which maps the admissible boundary currents to the far field pattern. In Chapter 1 we studied simple antenna models and formulated the optimization problems in terms of far field operators. As we have seen in Chapter 2, formulas (2.31a), (2.31b) and (2.72a)–(2.72c), far field patterns are intrinsically connected with the radiation of electromagnetic waves. Every change of the parameters of the system will evidently lead to a change of the far field. This dependence is what we model by the operator $\mathcal{K}$.

We will restrict ourselves to the case where the geometry, the spacing, the wave number, and other constitutive parameters are fixed and only the current ψ is at our disposal. In most of these cases, the far field $\mathcal{K}\psi$ depends linearly on ψ i.e., the **far field operator** $\mathcal{K}$ is a *linear* operator defined on some linear space. In some important cases as e.g., the finite array case, this operator is explicitly given. However, in many other cases only some properties of this operator can be derived and the operator itself has to be computed numerically. Some examples will be given below. To cover a variety of practical situations with one theory we take $\mathcal{K}$ to be simply a linear operator from some Hilbert space X over the field $\mathbb{C}$ of complex numbers into the space $C(S^{d-1})$ of continuous functions on the unit sphere in $\mathbb{R}^3$ (if $d = 3$) or the unit circle in $\mathbb{R}^2$ (if $d = 2$). We call $\mathcal{K}\psi$ the **far field pattern** corresponding to ψ. As typical examples for the space X we can take subspaces of the space $C(\Gamma)$ of continuous functions on some compact set or, if we allow jumps, subspaces of $L^2(\Gamma)$. Again, we denote the elements of the space X by ψ. Analogously, without distinction in notation, $C(S^{d-1})$ could be either the space of continuous *scalar* functions $f : \mathbb{R}^d \supset S^{d-1} \longrightarrow \mathbb{C}$ or be a space of continuous **tangential vector fields** $f : \mathbb{R}^d \supset S^{d-1} \longrightarrow \mathbb{C}^d$, i.e. satisfy $f(\hat{x}) \cdot \hat{x} = 0$ for all $\hat{x} \in S^{d-1} \subset \mathbb{R}^d$. In either case, $C(S^{d-1})$ is a Banach space with respect to the norm

$$\|f\|_{C(S^{d-1})} = \max_{\hat{x} \in S^{d-1}} |f(\hat{x})|\,, \quad f \in C(S^{d-1}) \tag{3.33}$$

where $|a|$ denotes the complex modulus if $a \in \mathbb{C}$ (a complex number) or $|a|$ denotes the vector norm $|a| = \sqrt{\sum_{j=1}^{N} |a_j|^2}$ if $a \in \mathbb{C}^N$ (a complex vector). Notice that we will not distinguish typographically between scalar and vector far fields except in specific examples, as the nature of the far field will be evident from the context.

We assume, furthermore, that the far field operator

$$\mathcal{K} : X \longrightarrow C(S^{d-1})$$

is also **compact**. This means that the image $\{\mathcal{K}\psi_n\}_{n=1}^{\infty} \subset C(S^{d-1})$ of every bounded sequence $\{\psi_n\}_{n=1}^{\infty} \subset X$ contains a subsequence which converges uniformly to some continuous function on S^{d-1}, see Section A.5 of the Appendix.

As a first example we refer to the case of an array of $2N+1$ elements at locations $\boldsymbol{y}_n \in \mathbb{R}^3$, $n = -N, \dots, N$, as discussed in Chapter 1. In this case, $\mathcal{K}$ is given by

$$(\mathcal{K}\boldsymbol{a})(\hat{\boldsymbol{x}}) = \sum_{n=-N}^{N} a_n \, \mathrm{e}^{-ik\boldsymbol{y}_n \cdot \hat{\boldsymbol{x}}}, \quad \hat{\boldsymbol{x}} \in S^2 \subset \mathbb{R}^3, \; \boldsymbol{a} = (a_{-N}, \dots, a_N)^\top \in \mathbb{C}^{2N+1}. \tag{3.34}$$

We note that, for this example, the dependence of $\mathcal{K}$ on the parameters k and $\boldsymbol{y}_n$ is explicit. Here we choose X to be the finite dimensional space $\mathbb{C}^{2N+1}$, equipped with the usual Euclidean norm, and note that $\mathcal{K}\boldsymbol{a} \in C(S^2)$ is a scalar function. This operator $\mathcal{K} : \mathbb{C}^{2N+1} \longrightarrow C(S^2)$ is certainly compact since X is finite dimensional.

As a second example we refer to the case of a line source of *arbitrary* shape. In Subsections 1.5.1 and 1.5.2 we considered linear line sources and circular line sources, respectively. Their far field patterns have the general form

$$\boldsymbol{E}_\infty(\hat{\boldsymbol{x}}) = \left|\hat{\boldsymbol{x}} \times (\hat{\boldsymbol{p}} \times \hat{\boldsymbol{x}})\right| \int_C \psi(\boldsymbol{y}) \, \mathrm{e}^{-ik\hat{\boldsymbol{x}}\cdot\boldsymbol{y}} \, dS(\boldsymbol{y}), \quad \hat{\boldsymbol{x}} \in S^2, \tag{3.35}$$

where $C \in \mathbb{R}^3$ denotes the shape of the curve. The part

$$f(\hat{\boldsymbol{x}}) = \int_C \psi(\boldsymbol{y}) \, \mathrm{e}^{-ik\hat{\boldsymbol{x}}\cdot\boldsymbol{y}} \, dS(\boldsymbol{y}), \quad \hat{\boldsymbol{x}} \in S^2,$$

is the line factor. In this example the operator $\mathcal{K}$ can be either defined as

$$(\mathcal{K}\psi)(\hat{\boldsymbol{x}}) = f(\hat{\boldsymbol{x}}) = \int_C \psi(\boldsymbol{y}) \, \mathrm{e}^{-ik\hat{\boldsymbol{x}}\cdot\boldsymbol{y}} \, dS(\boldsymbol{y}), \quad \hat{\boldsymbol{x}} \in S^2, \tag{3.36a}$$

or as

$$(\mathcal{K}\psi)(\hat{\boldsymbol{x}}) = \left|\hat{\boldsymbol{x}}\times(\hat{\boldsymbol{p}}\times\hat{\boldsymbol{x}})\right| f(\hat{\boldsymbol{x}}) = \left|\hat{\boldsymbol{x}}\times(\hat{\boldsymbol{p}}\times\hat{\boldsymbol{x}})\right| \int_C \psi(\boldsymbol{y}) \, \mathrm{e}^{-ik\hat{\boldsymbol{x}}\cdot\boldsymbol{y}} \, dS(\boldsymbol{y}), \tag{3.36b}$$

$\hat{x} \in S^2$, depending on whether one wants to formulate the quantities as directivity or signal-to-noise ratio (see Section 3.4) in terms of the line factor or the far field. In any case, the operator $\mathcal{K}$ is also given explicitly but through an integral instead of a finite sum. Here, X must be an infinite dimensional function space as, e.g., a subspace of $C(\Gamma)$ or of $L^2(\Gamma)$ or of some Sobolev space. In all these settings, the operator $\mathcal{K}$ given by (3.36a) or (3.36b) is compact.

As a third, and more general, case recall that we introduced the notion of the far field $\boldsymbol{E}_\infty$ associated with any radiating solution of Maxwell's equations (see § 2.9). In the cases where the generating sources lie outside the antenna as e.g., for the slot or reflector antennas, the components of the electromagnetic field must satisfy certain boundary conditions at the surface of the antenna. Typical boundary conditions were discussed previously in Section 2.11. In these cases, the currents ψ enter at the right hand sides of the boundary conditions. The fact that the corresponding boundary value problems have unique solutions allows us, once again, to write the far field $\boldsymbol{E}_\infty$ symbolically as $\mathcal{K}\psi$. In these cases, however, the operator $\mathcal{K}$ is not known explicitly but is only described implicitly through, e.g., boundary integral equations (see Section 5.3 or Section 6.2). Nevertheless, our analysis of the far field patterns in Chapters 5 and 6 will show that $\boldsymbol{E}_\infty$ is analytic, a fact which implies that the operator $\mathcal{K}$ from X into $C(S^{d-1})$ is well defined. However, we still have to specify concretely the space X of admissible input currents, and this choice is dictated by the exterior boundary value problem or, more exactly, our proof of the existence and uniqueness theorem (see Theorem 6.8) for that problem. Indeed, the existence of the operator $\mathcal{K}$ can be established only by such an existence and uniqueness proof, and its properties depend on the choice of space of boundary values. For all relevant cases, however, the operator $\mathcal{K}$ is not only bounded but even compact.

3.4 Measures of Antenna Performance

Having looked carefully at the properties of the far field operator $\mathcal{K}$ in the preceding section, we turn to the description of particular functionals that are traditionally of interest in the mathematical theory of antenna optimization. Our object is to specify the form of these functionals all of which involve the far field and to determine which of their properties as, for example, continuity or sequential upper-semicontinuity, properties that were discussed earlier in this chapter, are relevant to the various optimization problems.

Traditional measures of antenna performance involve a number of scalar quantities including *directivity*, *gain*, and *signal-to-noise ratio*. Other criteria may also be useful as, for example, in the classical Dolph-Tschebyscheff problem, the beam width and the side-lobe level. We have given precise definitions in Chapter 1 for the case of arrays and line sources. In this section, we will consider the analogous criteria for more general structures. They are usefully expressed in terms of the quantities that we have introduced in Chapter 2, namely far field patterns $\boldsymbol{E}_\infty$, surface currents, and power P_∞. In general it is not possible to express the far field pattern as a product of a single term representing a "reference source" and an "array factor". It is for this reason that we use the simple notation f to represent the actual far field $\boldsymbol{E}_\infty$. No problem occurs if one changes f to be a "factor" only, i.e. differs from $\boldsymbol{E}_\infty$ by just a multiplicative factor which is independent of the feeding. Once again we write ψ for the current.

We begin with the quantity called *directivity* (compare with the Definition 1.1).

Definition 3.29. *The* **directivity** *of an antenna in a given direction* $\hat{\boldsymbol{x}} \in S^2$ *is, as in (1.20),*

$$D_f(\hat{\boldsymbol{x}}) \;:=\; \frac{|f(\hat{\boldsymbol{x}})|^2}{\frac{1}{4\pi}\int_{S^2} |f(\hat{\boldsymbol{x}}')|^2\, dS}\,, \quad \hat{\boldsymbol{x}} \in S^2\,. \tag{3.37}$$

The quantity $\frac{1}{4\pi}\int_{S^2} |f(\hat{\boldsymbol{x}}')|^2\, dS$ is the total power radiated into the far field, see Section 2.9.

Using the representation (3.37) and the relation $f = \mathcal{K}\psi$, the directivity $D_f(\hat{\boldsymbol{x}})$ may also be written as

$$D_\psi(\hat{\boldsymbol{x}}) \;=\; 4\pi\, \frac{|(\mathcal{K}\psi)(\hat{\boldsymbol{x}})|^2}{\|\mathcal{K}\psi\|^2_{L^2(S^2)}}\,. \tag{3.38}$$

Here, we indicate the feeding, ψ, instead of the far field pattern, f.

The **gain** of an antenna, measured with respect to a given direction is usually defined as the ratio of the power radiated in that direction to the power field fed to the antenna. As there is always some dissipative loss and we are ignoring the questions of the efficiency of the feeding mechanism by which power fed to the antenna is converted into surface current ψ, we will use the term **radiation efficiency** (see (1.54a)) for the quantity

$$G_\psi(\hat{\boldsymbol{x}}) \;=\; \frac{|f(\hat{\boldsymbol{x}})|^2}{\|\psi\|^2_X} \;=\; \frac{|(\mathcal{K}\psi)(\hat{\boldsymbol{x}})|^2}{\|\psi\|^2_X}\,, \quad \hat{\boldsymbol{x}} \in S^2\,, \tag{3.39}$$

and the corresponding maximal radiation efficiency $\max\{G_\psi(\hat{\boldsymbol{x}}) : \hat{\boldsymbol{x}} \in S^2\}$. In this context we consider $\|\psi\|^2_X$ as a measure of the power fed to the antenna. The quantity $G_\psi(\hat{\boldsymbol{x}})$ coincides with the usual concept of gain only if all the power fed to the antenna were converted to surface current.

Likewise, the concept of **quality factor** or, briefly, Q-factor, which has been introduced by various authors in various ways (see e.g., [115]), will be defined in this book as

$$Q_\psi \;:=\; \frac{\|\psi\|^2_X}{\|f\|^2_{L^2(S^2)}} \;=\; \frac{\|\psi\|^2_X}{\|\mathcal{K}\psi\|^2_{L^2(S^2)}}\,. \tag{3.40}$$

This definition is connected with the far field operator $\mathcal{K}$ in a fundamental way. Specifically, we can compute the norm of the operator $\mathcal{K}$ considered as a mapping between the Hilbert spaces X and $L^2(S^2)$ and find

$$\|\mathcal{K}\|^2 = \sup_{\psi \in X} \frac{\|\mathcal{K}\psi\|^2_{L^2(S^2)}}{\|\psi\|^2_X} = \sup_{\psi \in X} \left(\frac{1}{Q_\psi} \right) \tag{3.41}$$

and hence

$$\inf_{\psi \in X} Q_\psi = \frac{1}{\|\mathcal{K}\|^2}. \tag{3.42}$$

The notions of directivity and radiation efficiency as given above by (3.37) and (3.39), respectively, represent an idealization of quantities that can actually be measured. It is more realistic to interpret measurements of the intensity in the far field as averages over (perhaps small) patches of the unit sphere. In particular, let $\alpha(\hat{\boldsymbol{x}})$ denote the characteristic function of a measurable sector $\mathcal{A}$ of the unit sphere S^2 i.e.,

$$\alpha(\hat{\boldsymbol{x}}) = \begin{cases} 1, & \hat{\boldsymbol{x}} \in \mathcal{A}, \\ 0, & \hat{\boldsymbol{x}} \notin \mathcal{A}. \end{cases} \tag{3.43}$$

Then we may generalize the concepts of directivity and radiation efficiency in a particular direction by replacing the expression $|f(\hat{\boldsymbol{x}})|^2$ with an average over a sector containing the particular direction $\hat{\boldsymbol{x}}$:

$$\frac{1}{\|\alpha\|^2_{L^2(S^2)}} \int_{S^2} \alpha(\hat{\boldsymbol{x}})^2 \, |f(\hat{\boldsymbol{x}})|^2 \, dS(\hat{\boldsymbol{x}}) = \frac{\|\alpha f\|^2_{L^2(S^2)}}{\|\alpha\|^2_{L^2(S^2)}}.$$

Then

$$D_{\psi,\alpha} = 4\pi \frac{\|\alpha(\mathcal{K}\psi)\|^2_{L^2(S^2)}}{\|\alpha\|^2_{L^2(S^2)} \, \|\mathcal{K}\psi\|^2_{L^2(S^2)}}, \tag{3.44}$$

and

$$G_{\psi,\alpha} = \frac{\|\alpha f\|^2_{L^2(S^2)}}{\|\alpha\|^2_{L^2(S^2)} \, \|\psi\|^2_X} = \frac{\|\alpha(\mathcal{K}\psi)\|^2_{L^2(S^2)}}{\|\alpha\|^2_{L^2(S^2)} \, \|\psi\|^2_X}, \tag{3.45}$$

are, respectively, the **generalized directivity** and **radiation efficiency** in the sector characterized by α. We drop the index α in the notation when there is no chance of confusion. We remark that, if the sector is the entire unit sphere, then

$$D_{\psi,\alpha} = 1 \quad \text{and} \quad G_{\psi,\alpha} = \frac{1}{4\pi Q_\psi}. \tag{3.46}$$

Thus the problem of maximizing the reciprocal of Q is a special case of maximizing the radiation efficiency. If we define the operator $\alpha\mathcal{K} : X \longrightarrow L^2(S^2)$ by

$$(\alpha\mathcal{K})\psi(\hat{\boldsymbol{x}}) := \alpha(\hat{\boldsymbol{x}})(\mathcal{K}\psi)(\hat{\boldsymbol{x}}), \quad \hat{\boldsymbol{x}} \in S^2, \psi \in X, \tag{3.47}$$

then the maximum radiation efficiency is given by

$$\sup_{\psi\in X} G_{\psi,\alpha} = \frac{\|\alpha\mathcal{K}\|^2}{\|\alpha\|^2_{L^2(S^2)}}. \tag{3.48}$$

These definitions and examples will give the reader an idea of some of the possible intrinsic criteria that can be used to measure the performance of antennas. To these we may add other functionals which are commonly used in the problem of **pattern synthesis**. Thus, for example, if a particular far field pattern f is given, we may be asked to find the inputs which minimize the mean-square deviation from this pattern in which case the functional has the form

$$\int\limits_{S^2} |(\mathcal{K}\psi)(\hat{\boldsymbol{x}}) - f(\hat{\boldsymbol{x}})|^2 \, dS, \quad \psi \in X. \tag{3.49}$$

We will consider the problem of pattern syntheses in a separate section. The following table shows the variety of such performance criteria or **cost functionals** that commonly occur, most of which we have already discussed.

Performance Criterion	Optimization Problem
Pattern matching to desired f (continuous)	min. $\mathcal{J}_1(\psi) = \int_{S^2} \|\mathcal{K}\psi - f\|^2 dS$,
Pattern matching to desired f (discrete)	min. $\mathcal{J}_2(\psi) = \sum_{i=1}^N \|(\mathcal{K}\psi)(\hat{\boldsymbol{x}}_i) - f(\hat{\boldsymbol{x}}_i)\|^2$,
Power in a sector with weight α	max. $\mathcal{J}_3(\psi) = \int_{S^2} \alpha^2 \|\mathcal{K}\psi\|^2 dS$,
Signal to noise ratio	max. $\mathcal{J}_4(\psi) = \dfrac{\|(\mathcal{K}\psi)(\hat{\boldsymbol{x}})\|^2}{\int_{S^2} \omega^2 \|\mathcal{K}\psi\|^2 dS}$,
Generalized signal to noise ratio	max. $\mathcal{J}_5(\psi) = \dfrac{\int_{S^2} \alpha^2 \|\mathcal{K}\psi\|^2 ds}{\int_{S^2} \alpha^2 \, dS \int_{S^2} \omega^2 \|\mathcal{K}\psi\|^2 dS}$,
Quality factor	min. $\mathcal{J}_6(\psi) = \dfrac{\\|\psi\\|^2_X}{\int_{S^2} \|\mathcal{K}\psi\|^2 dS}$,
Radiation efficiency	max. $\mathcal{J}_7(\psi) = \dfrac{\|(\mathcal{K}\psi)(\hat{\boldsymbol{x}})\|^2}{\\|\psi\\|^2_X}$,

Performance Criterion	Optimization Problem
Generalized radiation efficiency	max. $\mathcal{J}_8(\psi) = \dfrac{\displaystyle\int_{S^2} \alpha^2 \lvert\mathcal{K}\psi\rvert^2 dS}{\displaystyle\int_{S^2} \alpha^2\, dS\, \lVert\psi\rVert_X^2}$,
Directivity	max. $\dfrac{\lvert(\mathcal{K}\psi)(\hat{\boldsymbol{x}})\rvert^2}{\displaystyle\int_{S^2} \lvert\mathcal{K}\psi\rvert^2 dS}$,
Generalized directivity	max. $\dfrac{\displaystyle\int_{S^2} \alpha^2 \lvert\mathcal{K}\psi\rvert^2 dS}{\displaystyle\int_{S^2} \alpha^2\, dS \int_{S^2} \lvert\mathcal{K}\psi\rvert^2 dS}$

As we have seen in Section 3.2, for optimization problems the most important property of these functionals is their continuity with respect to the weak topology:

Theorem 3.30. *Let X be a Hilbert space, $\mathcal{K} : X \longrightarrow C(S^2)$ be compact, $f \in C(S^2)$, $\alpha, \omega \in L^\infty(S^2)$, and $\hat{\boldsymbol{x}}, \hat{\boldsymbol{x}}_i \in S^2$, $i = 1, \ldots, N$. Again let $\lVert\cdot\rVert_X$ be the norm on X and, for measurable and essentially bounded functions on S^2, let $\lVert\cdot\rVert$ be either the L^2-norm or the $L^\infty-$norm in $L^\infty(S^2)$.*

(a) The following functionals are weakly sequentially continuous on X:

$$\mathcal{J}_1(\psi) = \lVert\alpha(\mathcal{K}\psi - f)\rVert^2, \qquad \mathcal{J}_2(\psi) = \sum_{i=1}^{N} w_i\, \lvert(\mathcal{K}\psi)(\hat{\boldsymbol{x}}_i) - f(\hat{\boldsymbol{x}}_i)\rvert^2, \tag{3.50a}$$

$$\mathcal{J}_3(\psi) = \lVert\alpha\mathcal{K}\psi\rVert^2, \quad \mathcal{J}_4(\psi) = \frac{\lvert(\mathcal{K}\psi)(\hat{\boldsymbol{x}})\rvert^2}{\lVert\omega\mathcal{K}\psi\rVert^2}, \quad \mathcal{J}_5(\psi) = \frac{\lVert\alpha\mathcal{K}\psi\rVert^2}{\lVert\omega\mathcal{K}\psi\rVert^2}, \tag{3.50b}$$

where $w_i > 0$ are given weight factors. For the latter two functionals $\mathcal{J}_4$ and $\mathcal{J}_5$ we assume, in addition, that the map $\psi \mapsto \omega\mathcal{K}\psi$ is one-to-one. Then $\mathcal{J}_4$ and $\mathcal{J}_5$ are well defined on $X \setminus \{0\}$.

(b) The following functionals are weakly sequentially lower semi-continuous on $X\setminus\{0\}$, i.e. $\liminf_{n\to\infty} \mathcal{J}(\psi_n) \geq \mathcal{J}(\psi)$ for every sequence $\{\psi_n\}_{n=1}^\infty$ converging weakly to ψ:

$$\mathcal{J}_6(\psi) = \frac{\lVert\psi\rVert_X^2}{\lVert\mathcal{K}\psi\rVert^2}, \qquad -\mathcal{J}_7(\psi) = -\frac{\lvert(\mathcal{K}\psi)(\hat{\boldsymbol{x}})\rvert^2}{\lVert\psi\rVert_X^2}, \qquad -\mathcal{J}_8(\psi) = -\frac{\lVert\alpha\mathcal{K}\psi\rVert^2}{\lVert\psi\rVert_X^2}. \tag{3.50c}$$

Again, for $\mathcal{J}_6$ we assume that $\mathcal{K}$ is one-to-one.

Proof: We note that weak convergence $\psi_n \rightharpoonup \psi$ in X implies norm convergence $\|\mathcal{K}\psi_n - \mathcal{K}\psi\|_{C(S^2)} \to 0$ by the compactness of $\mathcal{K}$. Furthermore, this implies also convergence with respect to the $L^2(S^2)$–norm since S^2 has finite surface area. Therefore, also $\|\alpha\mathcal{K}\psi_n - \alpha\mathcal{K}\psi\|$ and $\|\omega\mathcal{K}\psi_n - \omega\mathcal{K}\psi\|$ converge to zero since both α and ω are bounded. This proves part (a).

For part (b) let $\{\psi_n\}_{n=1}^{\infty}$ converge to ψ weakly in X. Then the numerators of $\mathcal{J}_7$ and $\mathcal{J}_8$ converge to $|(\mathcal{K}\psi)(\hat{\boldsymbol{x}})|^2$ and $\|\alpha\mathcal{K}\psi\|^2$, respectively. From Theorem 3.5 we note that $\liminf_{n\to\infty} \|\psi_n\|_X \geq \|\psi\|_X$ which proves part (b). □

All of the constraints on the surface current, ψ, described in Subsection 3.2.2 have been specified directly in terms of the functions ψ. These are not the only types of constraints which define appropriate sets of admissible functions however. Indeed, as we have seen previously, the general inequality constraint $g(\psi) \leq 0$ may involve the far field $\mathcal{K}\psi$ itself. Since one important requirement on the constraint set, U, is its compactness in an appropriate (often weak) topology, it is important that the functions g_i which define U be weakly sequentially lower semi-continuous. This means in particular that any of the functions $\mathcal{J}_1, \ldots, \mathcal{J}_6$ analyzed in Theorem 3.30 can be used to define the constraint set. We give an example.

Example 3.31. It is possible, for example, to ask to maximize the power in a sector while attempting to keep the far field pattern close to a desired preassigned pattern. The resulting problem is, in some sense, a hybrid of the directivity and synthesis problems. In this case we consider the constrained problem

$$\text{Maximize} \quad \mathcal{J}_3(\psi) \quad \text{subject to } \psi \in U\,, \tag{3.51}$$

where

$$U \;=\; \bigl\{\psi \in U_0 : \mathcal{J}_1(\psi) \leq M\bigr\}\,, \tag{3.52}$$

for some closed and convex set $U_0 \subset X$. Here, $\mathcal{J}_1$ and $\mathcal{J}_3$ refer to the functionals introduced in Theorem 3.30 i.e. $\mathcal{J}_1(\psi) = \|\mathcal{K}\psi - f\|^2_{L^2(S^2)}$ and $\mathcal{J}_3(\psi) = \|\alpha\mathcal{K}\psi\|^2_{L^2(S^2)}$. The convexity of the set U follows easily from the convexity of $\mathcal{J}_1$ (see Lemma 3.32 below). Theorem 3.30 shows that $\mathcal{J}_1$ is weakly sequentially continuous and so the constraint set U is weakly sequentially closed in X. Note, however, that the set U is not bounded unless U_0 is bounded. In order to ensure the existence of optimal solutions one has to add a further constraint in U_0.

Likewise, a bound on the quality factor can be modeled by

$$g(\psi) \;=\; \|\psi\|_X^2 \;-\; \gamma\,\|\mathcal{K}\psi\|^2 \tag{3.53}$$

which will result in the non-convex set $U = \bigl\{\psi \in X : \|\psi\|_X^2 - \gamma\,\|\mathcal{K}\psi\|^2 \leq 0\bigr\}$.

Also, pointwise constraints on the far field pattern can be important. For example if α is the characteristic function of the sector in which power is to be focused, we may impose a constraint on the far field of the form

$$\big|[1-\alpha(\hat{\boldsymbol{x}})]f(\hat{\boldsymbol{x}})\big| \;\leq\; M \quad \text{for all } \hat{\boldsymbol{x}} \in S^2\,, \tag{3.54}$$

where M is a preassigned constant. The reader should recognize that this type of constraint differs significantly from the above form in that it describes an infinite number of scalar constraints. Nevertheless, this constraint can be used to describe a class of admissible control functions through the definition of $g : X \longrightarrow \mathbb{R}$ by

$$g(\psi) \;:=\; \sup_{\hat{\boldsymbol{x}}\in S^2}\big|[1-\alpha(\hat{\boldsymbol{x}})](\mathcal{K}\psi)(\hat{\boldsymbol{x}})\big| \;-\; M\,. \tag{3.55}$$

From the compactness of $\mathcal{K}$ it is easily seen that this function g is weakly sequentially continuous and therefore also weakly sequentially lower semicontinuous. Furthermore, the function g is convex. Indeed, for any $\lambda \in (0,1)$ we have

$$\begin{aligned} g\big(\lambda\psi + (1-\lambda)\varphi\big) &= \sup_{\hat{\boldsymbol{x}}\in S^2}\big|(1-\alpha(\hat{\boldsymbol{x}}))\big[\lambda\mathcal{K}\psi)(1-\lambda)\mathcal{K}\varphi\big]\big| \;-\; M \\ &\leq \sup_{\hat{\boldsymbol{x}}\in S^2}|\lambda(1-\alpha(\hat{\boldsymbol{x}}))\mathcal{K}\psi| \;+\; \sup_{\hat{\boldsymbol{x}}\in S^2}|(1-\lambda)(1-\alpha(\hat{\boldsymbol{x}}))\mathcal{K}(\varphi)| \\ &\quad - \lambda M \;-\; (1-\lambda)M \\ &= \lambda\, g(\psi) \;+\; (1-\lambda)\, g(\varphi)\,. \end{aligned}$$

However, the function g is not Fréchet differentiable which makes it impossible to apply the Lagrange multiplier rule of Theorem 3.22. Although there exist versions of the Lagrange multiplier rule which require only weaker forms of differentiability as e.g., the existence of a Gateaux variation, a practical, and perhaps more satisfying approach, is to constrain the real and imaginary parts in the form

$$\big|[1-\alpha(\hat{\boldsymbol{x}})]\mathrm{Re}\, f(\hat{\boldsymbol{x}})\big| \;\leq\; \tilde{M} \quad\text{and}\quad \big|[1-\alpha(\hat{\boldsymbol{x}})]\mathrm{Im}\, f(\hat{\boldsymbol{x}})\big| \;\leq\; \tilde{M} \quad \text{for all } \hat{\boldsymbol{x}} \in S^2\,,$$

and use methods of linear programming with an infinite number of constraints (see, e.g., [111]).

In order to apply the results on extreme points (Subsection 3.2.3) and the Lagrange-multiplier rule (Theorem 3.22) one must require further properties of the functionals $\mathcal{J}_j$ which will in fact be satisfied in our applications. We assume that the operator $\mathcal{K} : X \longrightarrow C(S^2)$ has the following properties:

(A1) $\mathcal{K} : X \longrightarrow C(S^2)$ is compact and one-to-one. In particular, $\mathcal{K}$ is not the zero operator.

(A2) $\mathcal{K}\psi \in C(S^2)$ is an analytic function on S^2 for every $\psi \in X$.

Moreover, we will, take the function $\alpha \in L^\infty(S^2)$ to be a real-valued, nonnegative, and to have the property:

(A3) the support of α, i.e. the closed set

$$\mathcal{A} := \bigcap \{A' \subset S^2 : A' \text{ closed and } \alpha = 0 \text{ a.e. on } S^2 \setminus A'\}$$

contains an open set (relative to S^2).

These three assumptions imply that for $\psi \neq 0$ the analytic function $\mathcal{K}\psi$ cannot even vanish on the support of α. Then we have:

Lemma 3.32. *Let $\mathcal{K} : X \longrightarrow C(S^2)$ and $\alpha \in L^\infty(S^2)$ satisfy the assumptions (A1), (A2), and (A3), and let $f \in L^2(S^2)$.*

(a) The functional

$$\mathcal{J}_1(\psi) := \|\alpha\mathcal{K}\psi - f\|^2_{L^2(S^2)}, \quad \psi \in X,$$

is strictly convex and continuously Fréchet differentiable with gradient

$$\nabla\mathcal{J}_1(\psi) = 2\,\mathcal{K}^*\big[\alpha(\alpha\mathcal{K}\psi - f)\big], \quad \psi \in X,$$

where $\mathcal{K}^* : L^2(S^2) \longrightarrow X$ denotes the adjoint of the operator $\mathcal{K}$ considered as an operator from X into $L^2(S^2)$.

(b) The functional

$$\mathcal{J}_2(\psi) := |(\mathcal{K}\psi)(\hat{\boldsymbol{x}}) - \gamma|^2, \quad \psi \in X \text{ with } \gamma \in \mathbb{C},\ \hat{\boldsymbol{x}} \in S^2 \text{ fixed},$$

is convex and continuously Fréchet differentiable with gradient

$$\nabla\mathcal{J}_2(\psi) = 2\left[(\mathcal{K}\psi)(\hat{\boldsymbol{x}}) - \gamma\right] p$$

where $p \in X$ denotes the Riesz representation of the linear functional $\varphi \mapsto (\mathcal{K}\varphi)(\hat{\boldsymbol{x}})$, $\varphi \in X$, i.e. the unique element $p \in X$ with

$$(\mathcal{K}\varphi)(\hat{\boldsymbol{x}}) = (\varphi, p)_X \quad \text{for all } \varphi \in X$$

(see Theorem A.31 of the Appendix).

Proof: Let $\lambda \in (0,1)$ and $\psi_1, \psi_2 \in X$. By the binomial formula it is readily seen that

$$\begin{aligned}
\mathcal{J}_1\big(\lambda\psi_1 + (1-\lambda)\psi_2\big) &= \|\lambda(\alpha\mathcal{K}\psi_1 - f) + (1-\lambda)(\alpha\mathcal{K}\psi_2 - f)\|^2_{L^2(S^2)} \\
&= \lambda\,\|\alpha\mathcal{K}\psi_1 - f\|^2_{L^2(S^2)} + (1-\lambda)\,\|\alpha\mathcal{K}\psi_2 - f\|^2_{L^2(S^2)} \\
&\quad - \lambda(1-\lambda)\,\|\alpha\mathcal{K}(\psi_1 - \psi_2)\|^2_{L^2(S^2)} \\
&= \lambda\,\mathcal{J}_1(\psi_1) + (1-\lambda)\,\mathcal{J}_1(\psi_2) \\
&\quad - \lambda(1-\lambda)\,\|\alpha\mathcal{K}(\psi_1 - \psi_2)\|^2_{L^2(S^2)}\,.
\end{aligned}$$

From this it follows that $\mathcal{J}_1\big(\lambda\psi_1 + (1-\lambda)\psi_2\big) \leq \lambda\,\mathcal{J}_1(\psi_1) + (1-\lambda)\mathcal{J}_1(\psi_2)$ and that equality holds if and only if $\alpha\mathcal{K}(\psi_1 - \psi_2)$ vanishes almost everywhere on

S^2. Since the support of α contains an open subset this implies that $\mathcal{K}(\psi_1 - \psi_2)$ vanishes almost everywhere on an open subset of S^2 and thus, by the assumption (A2) and the unique continuation principle for analytic functions, $\mathcal{K}(\psi_1 - \psi_2) = 0$ on S^2. Finally, the injectivity of $\mathcal{K}$ yields that $\psi_1 = \psi_2$ which ends the proof that $\mathcal{J}_1$ is strictly convex.

The Fréchet derivative of the function $\mathcal{J}_1$ has been computed in Example A.49 (b) of the Appendix. This completes the proof of part (a) of the lemma.

The proof of convexity of $\mathcal{J}_2$ follows the same argument as for the convexity of $\mathcal{J}_1$. The computation of the gradient of $\mathcal{J}_2$ is similar. Indeed, following the same method of computation as in Example A.49, we arrive at the intermediate result that

$$\begin{aligned} |\mathcal{K}(\psi+\varphi)(\hat{\boldsymbol{x}}) - \gamma|^2 \; - \; |(\mathcal{K}\psi)(\hat{\boldsymbol{x}}) - \gamma|^2 &= 2\,\mathrm{Re}\left[\left((\mathcal{K}\psi)(\hat{\boldsymbol{x}}) - \gamma\right)\overline{(\mathcal{K}\varphi)(\hat{\boldsymbol{x}})}\right] \\ &\quad + \; |(\mathcal{K}\varphi)(\hat{\boldsymbol{x}})|^2 . \end{aligned}$$

The Riesz Representation Theorem A.31 yields existence of some unique $p \in X$ such that $(\mathcal{K}\varphi)(\hat{\boldsymbol{x}}) = (\varphi, p)_X$ for all $\varphi \in X$ and so we may write

$$|\mathcal{K}(\psi+\varphi)(\hat{\boldsymbol{x}}) - \gamma|^2 - |(\mathcal{K}\psi)(\hat{\boldsymbol{x}})|^2 \; = \; 2\,\mathrm{Re}\left(\left[(\mathcal{K}\psi)(\hat{\boldsymbol{x}}) - \gamma\right] p\,,\, \varphi\right)_X + |(\mathcal{K}\varphi)(\hat{\boldsymbol{x}})|^2 \,.$$

From this equation we deduce that the gradient of $\mathcal{J}_2$ is given by

$$\nabla \mathcal{J}_2(\psi) \; = \; 2\left[(\mathcal{K}\psi)(\hat{\boldsymbol{x}}) - \gamma\right] p\,. \qquad (3.56)$$

This completes the proof of part (b). □

4

The Synthesis Problem

4.1 Introductory Remarks

The classical antenna synthesis problem is usually posed in one of two ways. The first addresses the problem of finding an aperture distribution which produces a far field that duplicates, or at least approximates, a prescribed pattern. The second involves determining the location and feed characteristics of a finite array of elementary sources which produce a radiation pattern that realizes some general property. In the latter case, for example, one may be given only some general characteristics of the far field pattern that are of interest, say that the beam be very wide in a vertical direction while remaining very narrow in the horizontal direction. Our concern in this chapter is the first of these problems, that of **pattern matching**, which we will discuss within our general framework of optimization problems, and then illustrate with concrete examples. Again, we make no attempt to survey the literature which, on this particular problem, is vast. Although exceptions may be found, most of the investigations in the existing literature proceed by means of specific cases. We will give a more general approach.

The concept of *pattern matching* usually treats specific requirements which effectively prescribe the far field pattern as a function of direction. In analyzing the pattern matching problem we meet almost immediately, the question of **realizability**, that is, whether it is possible to duplicate the desired pattern with the available input currents. Moreover, it may be required to match *both* the phase and amplitude of the far field, in which case we are dealing with what is called the **field synthesis problem**, or, in other cases simply the amplitude, the so-called **power synthesis problem**. It is this former problem on which we will concentrate.

As in our earlier discussion, we consider the geometry and constitutive parameters of the antenna to be fixed, ignore the specific feeding mechanism, and frame the problem strictly in terms of the surface current $\psi \in X$ on the antenna structure and the resulting far field pattern $f \in C(S^{d-1})$. They are related to one another, as before, by a compact operator K, and the problem is essentially one of solving a **Fredholm integral equation of the first kind**

$$K\psi \;=\; f\,. \tag{4.1}$$

If the function f belongs to the range of the operator K, we may ask for an exact solution, or an approximation of an exact solution, while in the case in which f fails to belong to the range of K (the case more typically met in applied problems) we must confront the more fundamental issue of deciding what should be meant by a "solution" and, only then, of finding appropriate techniques for its resolution.

In even the most elementary examples the range of the far field operator consists of smooth functions, so that, for example, piecewise constant far fields are strictly unattainable. But even if two functions f_1 and f_2 are far field patterns of surface currents ψ_1 and ψ_2, respectively, and if $f_1 - f_2$ is small the corresponding difference $\psi_1 - \psi_2$ doesn't have to be small at all. The synthesis problem is thus an example from an important class of problems of mathematical physics called **ill-posed**. We will discuss the nature of such problems in the next section.

To illustrate the ideas we will consider the case of the linear line source (see Subsection 1.5.1). The operator is then $K : L^2(-\ell,+\ell) \longrightarrow L^2(-1,+1)$, given by

$$(K\psi)(t) \;:=\; \int\limits_{-\ell}^{\ell} \psi(s)\, \mathrm{e}^{-ikst}\, ds\,, \quad |t| \leq 1\,. \tag{4.2}$$

In Theorem 1.11 we have shown that K is one-to-one. It is obvious that the range $\mathcal{R}(K)$ consists of analytic functions. Furthermore, the range is dense in $L^2(-1,+1)$. This follows from the fact that the L^2-adjoint $K^* : L^2(-1,+1) \longrightarrow L^2(-\ell,+\ell)$ of K, which is given by

$$(K^*\varphi)(t) \;:=\; \int\limits_{-1}^{1} \varphi(s)\, \mathrm{e}^{ikst}\, ds\,, \quad |t| \leq \ell\,, \tag{4.3}$$

is one-to-one since it is again the Fourier transform of an L^2-function.

The ill-posed nature of the synthesis problem also introduces difficulties from the computational point of view, as we will see later, arising particularly from the way in which the data of the problem, by which we mean the desired pattern, is presented. If the far field is prescribed, not exactly, but only through a set of data points, then questions of numerical sensitivity will arise. In particular, we must know how errors in the specification of f will effect the solution ψ.

Typically, there will also be certain *physical constraints* which it will be necessary to impose on the aperture distributions ψ. For example, undesirable effects may be caused by highly oscillatory distributions and it is usually desirable to work in a setting in which such oscillations are restricted. These physically based demands are usually referred to as realizability conditions for the surface currents and will be represented in our framework, by requiring the aperture functions to lie in some appropriate subset U, of the input space X. We will refer to the constrained functions as "**admissible**". Under certain conditions, these realizability constraints will serve to regularize the originally ill-posed problem, while in others, they will dictate compromises that will have to be made between requirements for accuracy of the approximating functions and the demands of meeting such *a priori* constraints.

4.2 Remarks on Ill-Posed Problems

Our object in this section is to remind the reader of, rather than to instruct him on the nature and some of the issues related to problems which are called ill-posed in the sense of Hadamard [45] first delineated in 1902. For a relevant discussion of the details, we refer the reader to the recent monograph of Kirsch [67].

According to Hadamard, any reasonable mathematical formulation of a physical problem leading to an equation of the form

$$K\psi = f, \tag{4.4}$$

where $f \in Y$ and the operator $K : X \longrightarrow Y$ are given, X, Y normed linear spaces, which must satisfy the conditions:

(1) The range of the operator K coincides with Y (**solvability**);
(2) The operator K should be one-to-one on its domain, that is, $Ku = Kv$ implies that $u = v$ (**uniqueness**); and
(3) The operator K should have a continuous inverse K^{-1} defined on Y (**stability**).

A model with these properties was called **well-posed** by Hadamard. Moreover, he illustrated, using the Cauchy problem for Laplace's equation, a problem which did *not* have these properties. He gave the name **ill-posed** to this and to other problems which did not satisfy one or more of the conditions listed above. The ill-posed problems that we will discuss here are of the form (4.4) in which K is compact, as is usually the case in applied problems. In particular, the problems of greatest interest to us involve the compact far field operator $\mathcal{K}$. The operator in (4.2) is an example. Below, we will investigate this example in more detail.

Since the 1902 paper of Hadamard, considerable attention has been paid to ill-posed problems as they are particularly important in a number of applied areas including geophysical problems, systems identification problems, and inverse scattering problems. Particularly important is the fundamental work of

Tikhonov, Ivanov, Lavrentiev, and Morozov (see [67] for detailed references). Our specific interest is the ill-posed nature of the antenna synthesis problem.[1] As we will discover, properties of the the range of the far field operator will be crucial for its resolution.

Under these circumstances, we need to confront the question of just what we mean when we say that we "solve" an ill-posed problem. Any notion of solution will clearly have to be the result of some compromise which can be dictated either by the mathematics, or the practical setting, or both. Thus in the case of the synthesis problem, we cannot necessarily realize the desired pattern exactly. Mathematically, we can only approximate the desired solution; practically, while often the acceptable level of error can be met, that level is attained only with surface currents having undesirable characteristics which degrade the antenna performance. Below we discuss particular regularization schemes, including the method of quasi-solutions and the method of Tikhonov.

In order to go further, we must be more precise about the properties of the range of the operator K. The goal is to describe a general setting within which to treat the operator equation

$$K\psi = f. \tag{4.5}$$

Let X and Y be Hilbert spaces and let $K : X \longrightarrow Y$ be a bounded linear operator whose range $\mathcal{R}(K)$ is not closed. Typically, K is a compact operator with infinite-dimensional range. Four situations can arise:

(i) $\mathcal{R}(K)$ may be dense (so that the null-space $\mathcal{N}(K^*)$ of the adjoint K^* is $\{0\}$), and $f \in \mathcal{R}(K)$;

(ii) $\mathcal{R}(K)$ is dense, and $f \notin \mathcal{R}(K)$;

(iii) The closure of $\mathcal{R}(K)$ is a proper subspace of Y (so that $Y = \overline{\mathcal{R}(K)} \oplus \mathcal{R}(K)^\perp$) and $f \in \mathcal{R}(K) \oplus \mathcal{R}(K)^\perp$;

(iv) The closure of $\mathcal{R}(K)$ is a proper subspace of Y, and $f \notin \mathcal{R}(K) \oplus \mathcal{R}(K)^\perp$.

Certainly, in the first of these cases, equation (4.5) has a classical solution while in case (ii) we can only hope for an approximate solution. The linear line source is an example of this kind as we have seen above.

An elegant treatment of linear ill-posed problems can be formulated by the use of the so-called **singular value decomposition**. We assume that $K : X \longrightarrow Y$ is compact. Then the operator K^*K is compact, self-adjoint and nonnegative; hence there exists an orthonormal system of eigenfunctions $\{\psi_n\}$ in X corresponding to the *positive* eigenvalues μ_n^2, i.e. $K^*K\psi_n = \mu_n^2\psi_n$. Defining $\varphi_n := \mu_n^{-1}K\psi_n$ we arrive at

[1] The ill-posed nature of this problem was first recognized by Bouwkamp and De Bruijn[18] in 1945. Application of Tikhonov regularization was suggested by Dechamps and Cabayan[33].

$$K\psi_n \;=\; \mu_n\,\varphi_n \quad\text{and}\quad K^*\varphi_n \;=\; \mu_n\,\psi_n\,, \tag{4.6}$$

and it is easy to see that $\{\varphi_n\}$ forms an orthogonal system in Y. The positive numbers μ_n are called the **singular values** of K. The set $\{\psi_n;\varphi_n;\mu_n\,:\,n=1,2,\ldots\}$ is called a **singular system** for K. In terms of this singular system, we may write, for any $\psi\in X$,

$$\psi \;=\; \sum_{n=1}^{\infty}(\psi,\psi_n)\,\psi_n \;+\; P_{\mathcal{N}}\psi \quad\text{and}\quad K\psi \;=\; \sum_{n=1}^{\infty}\mu_n\,(\psi,\psi_n)\,\varphi_n \tag{4.7}$$

where $P_{\mathcal{N}}$ is the orthogonal projection onto the null-space $\mathcal{N}(K)$. Formula (4.7) is called the singular value decomposition of the element ψ. Because of the orthogonality of the ψ_n, we can write

$$\|\psi\|^2 \;=\; \sum_{n=1}^{\infty}\bigl|(\psi,\psi_n)\bigr|^2 \;+\; \|P_{\mathcal{N}}\psi\|^2\,. \tag{4.8}$$

If K is one-to-one then $P_{\mathcal{N}}\psi = 0$ and the set $\bigl\{\psi_n\,:\,n\in\mathbb{N}\bigr\}$ is a *complete orthonormal system* in X.

From these remarks we can deduce a solvability condition for equation (4.5) first expounded by Picard.

Theorem 4.1. *The equation (4.5) has a solution for a given $f\in Y$ if and only if*

$$f\in\mathcal{N}(K^*)^{\perp} \quad\textit{and}\quad \sum_{n=1}^{\infty}\frac{1}{\mu_n^2}\,\bigl|(f,\varphi_n)\bigr|^2 \;<\; \infty, \tag{4.9}$$

a the solution is given by

$$\psi \;=\; \sum_{n=1}^{\infty}\frac{1}{\mu_n}\,(f,\varphi_n)\,\psi_n \tag{4.10}$$

Proof: If ψ is a solution of (4.5) then, necessarily, $(f,\varphi) \;=\; (K\psi,\varphi) \;=\; (\psi,K^*\varphi)=0$ for all $\varphi\in\mathcal{N}(K^*)$. The definition of a singular system yields

$$\mu_n(\psi,\psi_n) \;=\; (\psi,K^*\varphi_n) \;=\; (K\psi,\varphi_n) \;=\; (f,\varphi_n), \quad n=1,2,\ldots,$$

and so (4.8) implies that

$$\sum_{n=1}^{\infty}\frac{1}{\mu_n^2}\,\bigl|(f,\varphi_n)\bigr|^2 \;=\; \sum_{n=1}^{\infty}\bigl|(\psi,\psi_n)\bigr|^2 \;\le\; \|\psi\|^2\,.$$

Conversely, if f is orthogonal to $\mathcal{N}(K^*)$ and the condition (4.9) is satisfied, then the series

$$\psi \;:=\; \sum_{n=1}^{\infty}\frac{1}{\mu_n}\,(f,\varphi_n)\,\psi_n\,,$$

converges in X. Applying the operator K to both sides of this last equation yields

$$K\psi \;=\; \sum_{n=1}^{\infty} \frac{1}{\mu_n}\,(f,\varphi_n)\,K\varphi_n \;=\; \sum_{n=1}^{\infty}(f,\varphi_n)\,\varphi_n\,.$$

Comparing this with the singular value decomposition

$$f \;=\; f_0 \;+\; \sum_{n=1}^{\infty}(f,\varphi_n)\,\varphi_n$$

for some $f_0 \in \mathcal{N}(K^*)$ proves the assertion since by assumption $f \in \mathcal{N}(K^*)^\perp$ and so $f_0 = 0$. □

Conditions (4.9) are known as **Picard's criterion.** If $\mathcal{N}(K^*) = \{0\}$, then the range of K is dense and one is simply left with the second condition in (4.9) as the criterion for solvability of (4.5). This can be considered as an abstract smoothness condition since, in classical Fourier theory, the smoothness of a function is equivalent to a fast decay of its Fourier coefficients (or Fourier transform). We illustrate theses ideas by returning to the example of the linear line source.

Example 4.2. In the case of a linear line source the operator $K : L^2(-\ell,\ell) \longrightarrow L^2(-1,1)$ and its adjoint $K^* : L^2(-1,1) \longrightarrow L^2(-\ell,\ell)$ are given by (4.2) and (4.3), respectively. We compute $K^*K\psi$ by simply changing the order of integration:

$$\begin{aligned}(K^*K\psi)(t) &= \int_{-1}^{1}\int_{-\ell}^{\ell} \psi(\tau)\,\mathrm{e}^{-ik\tau s}\,d\tau\,\mathrm{e}^{ikst}\,ds \;=\; \int_{-\ell}^{\ell}\psi(\tau)\int_{-1}^{1}\mathrm{e}^{iks(t-\tau)}\,ds\,d\tau \\ &= 2\int_{-\ell}^{\ell}\psi(\tau)\,\frac{\sin k(t-\tau)}{k(t-\tau)}\,d\tau \;=\; 2\int_{-\ell}^{\ell}\psi(\tau)\,\mathrm{sinc}\,[k(t-\tau)]\,d\tau\,,\ \ |t|\le\ell\,,\end{aligned}$$

where the sinc-function is defined by

$$\mathrm{sinc}\; s \;:=\; \begin{cases} (\sin s)/s\,, & s \neq 0\,,\\ 1\,, & s = 0\,. \end{cases}$$

A singular system of K consists of functions $\psi_n \in L^2(-\ell,+\ell)$, $\varphi_n \in L^2(-1,+1)$, and $\mu_n > 0$ with $K\psi_n = \mu_n\varphi_n$ and $K^*\varphi_n = \mu_n\psi_n$. Furthermore, $\{\psi_n : n \in \mathbb{N}\}$ and $\{\varphi_n : n \in \mathbb{N}\}$ are complete orthonormal systems in $L^2(-\ell,+\ell)$ and $L^2(-1,+1)$, respectively. The completeness follows from the injectivity of K and K^*. Although there is no explicit expression of either ψ_n or φ_n, we can derive interesting and important properties of these functions. First, we note that ψ_n has an extension to an analytic function on all of $\mathbb{R}$

by considering the eigenvalue equation $K^*K\psi_n = \mu_n^2\psi_n$ in all of $\mathbb{R}$, i.e. we extend $\psi_n(t)$ for $|t| > \ell$ by

$$\psi_n(t) := \frac{2}{\mu_n^2}\int_{-\ell}^{\ell} \psi_n(\tau)\operatorname{sinc}[k(t-\tau)]\,d\tau\,, \quad |t| > \ell\,. \tag{4.11}$$

Note that also $\psi_n \in L^2(\mathbb{R})$. In the following, our notation does not distinguish between the original function $\psi_n \in L^2(-\ell,+\ell)$ and its analytic extension $\psi_n \in L^2(\mathbb{R})$ given by (4.11). Furthermore, we define the auxiliary functions $S, \Psi_n \in L^2(\mathbb{R})$ by

$$S(t) := \operatorname{sinc}(kt) = \begin{cases} \sin(kt)/(kt)\,, & t \neq 0\,, \\ 1\,, & t = 0\,, \end{cases}$$

and

$$\Psi_n(t) := \begin{cases} \psi_n(t)\,, & |t| \leq \ell\,, \\ 0\,, & |t| > \ell\,, \end{cases}$$

and observe from (4.11) that ψ_n is the weighted convolution

$$\psi_n = \frac{2}{\mu_n^2}\,\Psi_n * S\,. \tag{4.12}$$

Now we consider $\varphi_n = \frac{1}{\mu_n} K\psi_n$ as a Fourier transform, i.e.

$$\varphi_n(t) = \frac{1}{\mu_n}\int_{-\ell}^{\ell} \psi_n(s)\,\mathrm{e}^{-ikst}\,ds = \frac{\sqrt{2\pi}}{\mu_n}\,\widehat{\Psi}_n(kt)\,.$$

From the orthogonality of φ_n we conclude that

$$\begin{aligned}\int_{-k}^{k} \widehat{\Psi}_n(s)\,\overline{\widehat{\Psi}_m(s)}\,ds &= k\int_{-1}^{1} \widehat{\Psi}_n(kt)\,\overline{\widehat{\Psi}_m(kt)}\,dt \\ &= \frac{k\mu_n\mu_m}{2\pi}\int_{-1}^{1} \varphi_n(t)\,\overline{\varphi_m(t)}\,dt = \frac{k\mu_n^2}{2\pi}\,\delta_{nm}\,. \end{aligned} \tag{4.13}$$

Now we show that the functions $\{\psi_n\}$ enjoy a **double orthogonality property**: They are orthogonal not only in $L^2(-\ell,+\ell)$ but also in $L^2(\mathbb{R})$. First, if we take the Fourier transform of the convolution (4.12) then

$$\hat{\psi}_n = \frac{2}{\mu_n^2}\sqrt{2\pi}\,\hat{S}\,\widehat{\Psi_n}\,,$$

so that, by Plancherel's theorem (see, e.g. [145], p. 153), we have

$$(\psi_n, \psi_m)_{L^2(\mathbb{R})} \;=\; \left(\hat{\psi}_n, \hat{\psi}_m\right)_{L^2(\mathbb{R})} \;=\; \frac{8\pi}{\mu_n^2 \mu_m^2} \int\limits_{-\infty}^{\infty} |\hat{S}(t)|^2 \, \widehat{\Psi}_n(t) \, \overline{\widehat{\Psi}_m(t)} \, dt \, . \qquad (4.14)$$

If we compute the Fourier transform of S,

$$\hat{S}(t) \;=\; \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{\infty} \frac{\sin(ks)}{ks} \, \mathrm{e}^{-ist} \, ds \;=\; \frac{1}{k} \sqrt{\frac{\pi}{2}} \begin{cases} 1 \,, & |t| < k \,, \\ 0 \,, & |t| > k \,, \end{cases}$$

we may conclude that

$$(\psi_n, \psi_m)_{L^2(\mathbb{R})} \;=\; \frac{4\pi^2}{\mu_n^2 \mu_m^2 k^2} \int\limits_{-k}^{k} \widehat{\Psi}_n(t) \, \overline{\widehat{\Psi}_m(t)} \, dt \, ,$$

i.e. with (4.13), that

$$(\psi_n, \psi_m)_{L^2(\mathbb{R})} \;=\; \frac{2\pi}{k \mu_n^2} \, \delta_{nm} \;=\; \frac{\lambda}{\mu_n^2} \, \delta_{nm} \qquad (4.15)$$

which proves the double orthogonality property.[2].

We note that $\|\psi_n\|^2_{L^2(\mathbb{R})}$ is just the super-gain ratio (1.63) of ψ_n. Indeed, from (4.15) we observe that

$$\gamma_\lambda(\psi_n) \;=\; \lambda \, \frac{\|\psi\|^2_{L^2(-\ell,+\ell)}}{\|K\psi\|^2_{L^2(-1,+1)}} \;=\; \frac{\lambda}{\mu_n^2} \;=\; \|\psi_n\|^2_{L^2(\mathbb{R})} \, .$$

Let us now return to the four different cases at the beginning of this section which characterize the range $\mathcal{R}(K)$ of the operator K.

If we use Theorem 4.1, we can then make the simple observation that illustrates ill-posed nature of the problem

$$K\psi \;=\; f$$

when K is compact and the range is infinite dimensional. Specifically, let $\psi \in X$ be the solution of the equation $K\psi = f$ for the "unperturbed" right hand side $f \in Y$, and assume that $\tilde{f} = f + \delta\varphi_k$ denotes a (particular) perturbed right hand side. Then $|\delta|$ represents the error in the data i.e., $|\delta| = \|f - \tilde{f}\|$. The solution $\tilde{\psi}$ of the perturbed problem $K\tilde{\psi} = \tilde{f}$, is given by

[2] The usefulness of double orthogonal systems for expansions of functions in terms of a common set of basis functions was studied carefully by Slepian and Pollak in a series of papers beginning with [124]. Rhodes [115] discusses their use in the synthesis problem and gives specific examples (see also Angell and Nashed [10])

$$\tilde{\psi} = \psi + \frac{\delta}{\mu_k}\psi_k$$

since $K\tilde{\psi} = K\psi + \frac{\delta}{\mu_k}K\psi_k = f + \delta\varphi_k = \tilde{f}$. Therefore, the error

$$\|\tilde{\psi} - \psi\| = \frac{|\delta|}{\mu_k}$$

can be made arbitrarily large since $\mu_k \to 0$ as k tends to infinity.

4.3 Regularization by Constraints

Let us now look at the cases when $f \notin \mathcal{R}(K)$ i.e., when the equation $K\psi = f$ is not solvable. In light of Picard's criterion (Theorem 4.1), this is the generic case when f represents any approximation of the exact right hand side. In this case one is tempted to look at the least square solution of the equation in the following sense:

Definition 4.3. *A function f is a* **least-squares solution** *or* **quasi-solution** *of the operator equation*

$$K\psi = f \tag{4.16}$$

provided

$$\inf\{\|Ku - f\| : u \in X\} = \|K\psi - f\| . \tag{4.17}$$

Closely associated with the notion of quasi-solution, is that of a **generalized inverse** $K^\dagger f$. Since

$$\|Ku - f\|^2 = \|Ku - P_{\mathcal{R}}f\|^2 + \|f - P_{\mathcal{R}}f\|^2 ,$$

where $P_{\mathcal{R}}$ is the orthogonal projector of Y onto $\overline{\mathcal{R}(K)}$, it is clear that (4.17) holds if and only if $f \in \mathcal{R}(K) + \mathcal{R}(K)^\perp$, which is a dense set in Y. For such f the set of all least-squares solutions of (4.17), denoted by $\mathcal{S}_f$, is a nonempty, closed, convex set (indeed $\mathcal{S}_f$ is the translate of $\mathcal{N}(K)$ by a fixed element of $\mathcal{S}_f$) and hence has a unique element of minimal norm, denoted by $K^\dagger f$. The generalized inverse $K^\dagger$ is thus an operator from $\mathcal{R}(K) + \mathcal{R}(K)^\perp$ into X.

In light of the following theorem, the use of this weaker concept of quasi-solution does not overcome the ill posedness of the problem.

Theorem 4.4. *Let $K : X \longrightarrow Y$ be bounded and let $f \in Y$. The set of quasi-solutions is characterized by the set of solutions of the* **normal equation**

$$K^*K\psi = K^*f \tag{4.18}$$

where again $K^* : Y \longrightarrow X$ denotes the adjoint of K.

Proof: Using the binomial theorem in the Hilbert space X we have that

$$\begin{aligned}\|Ku-f\|^2 \;-\; \|K\psi-f\|^2 &= 2\,\mathrm{Re}\left(K\psi-f, K(u-\psi)\right) \;+\; \|K(u-\psi)\|^2 \\ &= 2\,\mathrm{Re}\left(K^*(K\psi-f), u-\psi\right) \\ &\quad+\; \|K(u-\psi)\|^2 \end{aligned} \tag{4.19}$$

for all $u, \psi \in X$.

First let $\psi \in X$ be a quasi-solution. Then the left hand side is non-negative. Fixing $\varphi \in X$ and setting $u = \psi + \epsilon\varphi$ for some $\epsilon > 0$, we conclude from (4.19) that

$$2\epsilon\,\mathrm{Re}\left(K^*(K\psi-f),\varphi\right) \;+\; \epsilon^2\,\|K\varphi\|^2 \;\geq\; 0\,.$$

Dividing by ϵ and letting ϵ tend to zero yields

$$\mathrm{Re}\left(K^*(K\psi-f),\varphi\right) \;\geq\; 0\,.$$

This holds for every $\varphi \in X$. From this (4.18) follows by substituting $\varphi = -K^*(K\psi - f)$.

Conversely suppose that (4.18) holds. Then the first term on the right hand side of (4.19) vanishes and thus $\|Ku-f\|^2 - \|K\psi-f\|^2 \geq 0$. This proves the theorem. □

We see from this result that quasi-solutions for a compact operator K are again characterized by the solutions of an operator equation with compact operator K^*K. Therefore, even the quasi-solutions do not depend continuously on the right hand side.

The situation is different when we impose additional constraints on the optimization problem (4.17). Indeed, one of the fundamental observations of Tikhonov is that the restriction of the problem to a compact set insures that the problem becomes well-posed. The use of such a constraint may be viewed as the utilization of *a priori* information and has long been recognized to play a significant role in bringing about continuous dependence. The general result can be expressed as:

Theorem 4.5. *Let X and Y be separable Hilbert spaces and $K : X \longrightarrow Y$ be a linear, bounded operator which is one-to-one. Let $C \subset X$ be weakly sequentially compact. Then $K(C)$ is weakly sequentially compact and $K^{-1} : Y \supset K(C) \longrightarrow X$ is weakly sequentially continuous.*

Proof: The compactness of $K(C)$ is an immediate consequence of the fact that bounded linear operators map weakly convergent sequences into weakly convergent sequences.

Let $\varphi_n \in K(C)$ converge weakly to some $\varphi \in K(C)$ i.e., $\varphi_n \rightharpoonup \varphi$. Setting $\varphi_n = K\psi_n$ and $\varphi = K\psi$ we have $K\psi_n \rightharpoonup K\psi$. Since C is weakly sequentially compact we can extract a subsequence $\left\{\psi_{n_k}\right\}_{k=1}^{\infty}$ with $\psi_{n_k} \rightharpoonup \hat{\psi}$ for some $\hat{\psi} \in$

C. This yields $K\psi_{n_k} \rightharpoonup K\hat{\psi}$ and thus $K\hat{\psi} = K\psi$, i.e., $\hat{\psi} = \psi$. Moreover this argument holds for every weakly convergent subsequence and so we conclude that the sequence $\{\psi_n\}$ itself converges weakly to ψ which proves that $\psi_n = K^{-1}\varphi_n \rightharpoonup \psi = K^{-1}\varphi$. □

To see how this result is applied to optimization problems, let us first consider the simplest case:

$$\text{Minimize} \quad \|K\psi - f\| \quad \text{subject to} \quad \psi \in X \text{ and } \|\psi\| \leq M\,, \tag{4.20}$$

where M is some *a priori* given positive number. For this constrained minimization problem, which we call the method of **restricted quasi-solutions**, it is possible to give necessary and sufficient conditions for the existence of an optimal solution of (4.20).

Theorem 4.6. *Let K be a bounded operator on X. For any constant $M > 0$, the optimization problem (4.20) has a solution $\psi^o \in X$. Moreover, if the range of K is dense in Y and $\|f\| > \inf\{\|K\psi - f\| : \psi \in X,\ \|\psi\| \leq M\} > 0$, then ψ^o is an optimal solution if and only if $\|\psi^o\| = M$, and there exists a constant $\eta > 0$ such that*

$$\left(K\psi^o, K\varphi\right)_Y \;+\; \eta\left(\psi^o, \varphi\right)_X \;=\; \left(f, K\varphi\right)_Y \quad \textit{for all } \varphi \in X\,, \tag{4.21a}$$

i.e.,

$$K^*K\psi^o \;+\; \eta\,\psi^o \;=\; K^*f\,. \tag{4.21b}$$

Proof: Existence of an optimal solution follows directly from the general existence result, Theorem 3.1, since the ball of radius R is weakly compact by Theorem 3.7 and the functional $\psi \mapsto \|K\psi - f\|$ is weakly lower semicontinuous as a (norm-) continuous and convex functional (Theorem 3.5).

Now let $\|f\| > J$ and let ψ^o be any optimal solution. We write the constraint as the inequality $h(\psi) = \|\psi\|^2 - M^2 \leq 0$ and note that the Fréchet derivative $2\psi^o \neq 0$ since we have assumed that $J < \|f\|$. Then there is a Lagrange multiplier $\eta \geq 0$ with

$$K^*(K\psi^o - f) \;+\; \eta\,\psi^o \;=\; 0\,, \tag{4.22}$$

and $\eta\big(\|\psi^o\|^2 - M^2\big) = 0$. It remains to show that, in fact, $\eta > 0$.

Indeed, if $\eta = 0$ then equation (4.22) becomes $K^*(K\psi^o - f) = 0$, which implies that $K\psi^o - f = 0$ since we have assumed that the range of K is dense in Y. This contradicts the assumption that $J > 0$. So we see that necessarily $\eta > 0$.

To prove the sufficiency of this condition, let $\eta > 0$ and suppose that $\psi^o \in X$ with $\|\psi^o\| = M$ satisfies the equation (4.21). Let ψ be any element of X, with $\|\psi\| \leq M$. The binomial formula yields

$$\begin{aligned}
&\|K\psi - f\|^2 \;+\; \eta\,\|\psi\|^2 \;-\; \|K\psi^o - f\|^2 \;-\; \eta\,\|\psi^o\|^2 \\
&\;= 2\,\mathrm{Re}\,\big(K\psi^o - f, K(\psi - \psi^o)\big) + \|K(\psi - \psi^o)\|^2 + 2\eta\,\mathrm{Re}\,(\psi^o, \psi - \psi^o) \\
&\qquad + \eta\,\|\psi - \psi^o\|^2 \\
&\;= 2\,\mathrm{Re}\,\big(\underbrace{K^*(K\psi^o - f) + \eta\psi^o}_{=0}, \psi - \psi^o\big) \;+\; \|K(\psi - \psi^o)\|^2 \;+\; \eta\,\|K\psi - \psi^o\|^2 \\
&\;\ge 0\,,
\end{aligned}$$

from which we conclude that

$$\|K\psi - f\|^2 \;-\; \|K\psi^o - f\|^2 \;\ge\; \eta\big[\|\psi^o\|^2 \;-\; \|\psi\|^2\big] \;=\; \eta\big[M^2 - \|\psi\|^2\big] \;\ge\; 0\,,$$

and the proof is complete. □

From this theorem we may again clearly observe the ill-posedness of the problem. If we think of the right hand side g as being polluted, then, generically, the assumptions of the theorem are satisfied and the restricted quasi-solution satisfies $\|\psi^o\| = M$ which is as *inaccurate* as the measurements of g. However, by using this method we have retained stability of the solution in the *weak* topology.

Theorem 4.7. *Let $\{f_n\}_{n=1}^\infty \subset Y$ converge to f and let ψ_n^o with $\|\psi_n^o\| \le M$ be any restricted quasi-solution corresponding to f_n, $n = 1, 2 \ldots$. Then there exist weak accumulation points of the sequence $\{\psi_n^o\}_{n=1}^\infty$, and every such weak accumulation point ψ^* is optimal for f. Furthermore, the optimal values converge, i.e. $\|K\psi_n^o - f_n\| \to \|K\psi^* - f\|$.*

Proof: The existence of weak accumulation points follows again from Theorem A.58. Suppose that ψ^* is such an accumulation point. Then if the subsequence $\left\{\psi_{n_k}^o\right\}_{k=1}^\infty$ converges weakly to ψ^* we have

$$\|\psi^*\| \;\le\; \liminf_{k\to\infty} \left\|\psi_{n_k}^o\right\| \;\le\; M$$

and

$$K\psi_{n_k}^o - f_{n_k} \;\rightharpoonup\; K\psi^* - f\,.$$

Let ψ^o be any optimal solution corresponding to f. Then we have

$$\begin{aligned}
J \le \|K\psi^* - f\| \;&\le\; \liminf_{k\to\infty} \left\|K\psi_{n_k}^o - f_{n_k}\right\| \;\le\; \limsup_{k\to\infty} \left\|K\psi_{n_k}^o - f_{n_k}\right\| \\
&\le \lim_{k\to\infty} \|K\psi^o - f_{n_k}\| \;=\; \|K\psi^o - f\| \;=\; J\,,
\end{aligned}$$

which proves that ψ^* is optimal and that

$$\lim_{k\to\infty} \left\|K\psi_{n_k}^o - f_{n_k}\right\| \;=\; \|K\psi^* - f\| \;=\; J\,.$$

Since this equality holds for every subsequence it follows that

$$\lim_{n\to\infty} \|K\psi_n^o - f_n\| = J,$$

and the proof is complete. □

We note that a very similar result has been proven in already Section 3.3 (Theorem 3.25). However, here we consider *perturbations of the data f* while in Section 3.3 we study finite dimensional approximations.

Although it seems that we have not gained much by the introduction of the restricted quasi-solution, we point out that stability with respect to the weak topology of X can lead to a very strong result if we know *a priori* that the solution ψ is "smooth" i.e., contained in a subspace $X_1 \subset X$ such that the imbedding $j : X_1 \hookrightarrow X$ is compact. Instead of $K : X \longrightarrow Y$ we consider rather $K_1 := K \circ j : X_1 \longrightarrow Y$. Then weak convergence $\psi_n \rightharpoonup \psi$ in X_1 implies norm convergence $\psi_n \to \psi$ in X since the imbedding operator j is compact. Therefore, by replacing the constraint $\|\psi\|_X \leq M$ by a stronger one, $\|\psi\|_{X_1} \leq M$ we retain stability with respect to the norm in X. We note that the adjoint of K_1 is given by $K_1^* = j^* \circ K^*$ where $j^* : X \longrightarrow X_1$ denotes the adjoint of the imbedding j. Fortunately, in many cases it is not necessary to compute j^* explicitly. Indeed, the variational equation (4.21a) takes the form

$$\left(K\psi^o, K\varphi\right)_Y + \eta\left(\psi^o, \varphi\right)_{X_1} = \left(f, K\varphi\right)_Y \quad \text{for all } \varphi \in X_1$$

or, taking the adjoint $K^* : Y \longrightarrow X$,

$$\left(K^*K\psi^o, \varphi\right)_X + \eta\left(\psi^o, \varphi\right)_{X_1} = \left(K^*f, \varphi\right)_X \quad \text{for all } \varphi \in X_1 .$$

This equation can be exploited explicitly (see Example 4.11 below).

Let us return to a more concrete setting. In Chapter 1 we have introduced the **super-gain ratio** $\gamma_\lambda(\varphi)$ (see Definition 1.12 and (1.63)) and the **quality factor** Q (see Definition 1.9 and (1.54b)) of a line source which have the form $\|\psi\|^2 / \|K\psi\|^2$. In the following optimization problem we investigate constraints on the (abstract) super-gain ratio:

$$\text{Minimize} \quad \|K\psi - f\| \quad \text{subject to} \quad \|\psi\| \leq M\,\|K\psi\|\,, \tag{4.23}$$

where $M > 0$ is some given constraint. We assume that $M > 1/\,\|K\|$ in order to insure that non-trivial elements ψ satisfy the constraint. Indeed, by the definition of $\|K\| := \sup_{\psi\neq 0} \|K\psi\| \,/\, \|\psi\|$, we observe that there exists a $\psi \neq 0$ with $\|K\psi\| \,/\, \|\psi\| \geq 1/M$ provided $1 < M\,\|K\|$.

First we show existence of the optimization problem (4.23).

Theorem 4.8. *Let $K : X \longrightarrow Y$ be compact and $M\,\|K\| > 1$. Then there exists a solution ψ^o of the constrained optimization problem (4.23).*

Proof: Let $R := M\big(2\,\|f\| + 1\big)$, and define the set U by

$$U \;:=\; \big\{\psi \in X : \|\psi\| \leq R,\; \|\psi\| \leq M\,\|K\psi\|\big\}.$$

For $\psi \notin U$ with $\|\psi\| \leq M\,\|K\psi\|$, we have that $\|\psi\| > R$ and thus

$$\|K\psi - f\| \;\geq\; \|K\psi\| - \|f\| \;\geq\; \frac{1}{M}\,\|\psi\| - \|f\| \;>\; \frac{R}{M} - \|f\| \;=\; \|f\| + 1,$$

so that ψ *cannot* be optimal ($\psi = 0$ has a smaller defect). The set U is clearly nonempty, bounded and in fact weakly sequentially compact. To see this, it suffices to show that this set is weakly closed. To this end, let $\{\psi_n\}_{n=1}^\infty \subset U$, and suppose $\psi_n \rightharpoonup \psi$. Then

$$\|\psi\| \;\leq\; \liminf_{n\to\infty} \|\psi_n\| \;\leq\; M \liminf_{n\to\infty} \|K\psi_n\| \;=\; M\,\|K\psi\|$$

since $K\psi_n \to K\psi$ by the compactness of K.

Furthermore, we note that the objective function is weakly continuous since K is compact. Therefore, Theorem 3.1 is applicable and yields existence of a solution $\psi^o \in X$ of the optimization problem which is, in fact, in U. This completes the proof. □

If we now wish to apply the multiplier rule again we must compute the Fréchet derivative of the constraint $h(\psi) = \|\psi\|^2 - M^2\,\|K\psi\|^2$, which in fact is $h'(\psi) = 2\big[\psi - M^2 K^* K\psi\big]$.

Theorem 4.9. *Let ψ^o be some optimal solution of (4.23). Let the following* **constraint qualification** *be satisfied:*

$$\psi^o \;-\; M^2 K^* K \psi^o \;\neq\; 0, \tag{4.24}$$

i.e. that ψ^o is a so-called regular point. Then there exists $\eta \geq 0$ such that $\eta\big(\|\psi^o\|^2 - M^2\,\|K\psi^o\|^2\big) = 0$ and

$$K^* K\psi^o \;+\; \eta\,\big(\psi^o - M^2\,K^* K\psi^o\big) \;=\; K^* f. \tag{4.25}$$

Under the additional assumption that K^* is one-to-one and $K\psi^o \neq f$, the Lagrange multiplier η is strictly positive and $\|\psi^o\| = M\,\|K\psi^o\|$.

We remark that (4.24) is certainly satisfied if $1/M^2$ is not an eigenvalue of K^*K.

We now consider *perturbations* of this optimization problem. Let $\{f_n\}_{n=1}^\infty \subset Y$ be a sequence with $f_n \to f$ as $n \to \infty$, and for any integer $n = 1, 2, \ldots$ let $\psi_n^o \in X$ be an optimal solution corresponding to f_n. Making the assumptions that

(i) K^* is one-to-one;

(ii) $K\psi_n^o \neq f_n$ for $n = 1, 2, \ldots$, and
(iii) $\psi_n^o - M^2 K^* K \psi_n^o \neq 0$ for all n,

then application of Theorem 4.9 asserts that

$$\|\psi_n^o\| \;=\; M\;\|K\psi_n^o\|\,, \quad n = 1, 2, \ldots$$

Moreover, the sequence $\{\psi_n^o\}_{n=1}^\infty$ is bounded since we have

$$\begin{aligned}\|\psi_n^o\| = M\,\|K\psi_n^o\| \;&\leq\; M\big(\|K\psi_n^o - f_n\| + \|f_n\|\big) \\ &\leq M\big(\|K0 - f_n\| + \|f_n\|\big) \;=\; 2M\,\|f_n\|\,.\end{aligned}$$

Since the sequence is bounded, there exist weak accumulation points. Suppose that ψ^* is one such accumulation point and that $\psi_{n_k} \rightharpoonup \psi^*$. Then, again, $K\psi_{n_k}^o \to K\psi^*$ and thus, for some optimal ψ^o corresponding to f,

$$\|\psi^*\| \;\leq\; \liminf_k \big\|\psi_{n_k}^o\big\| \;=\; M \lim_{k\to\infty} \big\|K\psi_{n_k}^o\big\| \;=\; M\;\|K\psi^*\|\,.$$

Furthermore,

$$\begin{aligned}J \leq \|K\psi^* - f\| \;&=\; \lim_{k\to\infty} \big\|K\psi_{n_k}^o - f_{n_k}\big\| \;\leq\; \lim_{k\to\infty} \|K\psi^o - f_{n_k}\| \\ = \|K\psi^o - f\| \;&=\; J\,,\end{aligned}$$

that is

$$\|K\psi^* - f\| \;=\; \lim_{k\to\infty} \|K\psi_{n_k} - f_{n_k}\| \;=\; J\,.$$

In particular, ψ^* is optimal and the optimal values converge to the optimal value corresponding to f.

If, for the weak accumulation point ψ^*, the constraint qualification (4.24) is satisfied, K^* is one-to-one and $K\psi^* \neq f$, then we have even convergence in norm since then $\|\psi^*\| = M\,\|K\psi^*\|$ and so

$$\big\|\psi_{n_k}^o\big\| \;=\; M\,\big\|K\psi_{n_k}^o\big\| \;\longrightarrow\; M\,\|K\psi^*\| \;=\; \|\psi^*\|\,.$$

From this and the weak convergence, we conclude that

$$\big\|\psi_{n_k}^o - \psi^*\big\|^2 \;=\; 2\,\mathrm{Re}\,\big(\psi^* - \psi_{n_k}^o, \psi^*\big) \;-\; \|\psi^*\|^2 \;+\; \big\|\psi_{n_k}^o\big\|^2 \;\longrightarrow\; 0$$

that is, we have norm convergence of the $\psi_{n_k}^o$ to ψ^*.

4.4 The Tikhonov Regularization

The contribution of Tikhonov to the treatment of ill-posed problems, usually called a "regularization" procedure, was to show that the addition of a non-negative functional to the original functional $\|K\psi - f\|$ will stabilize the ill-posed problem. This contrasts with the method of the preceeding section which entails the adding of an *a priori* requirement that the optimal solution belong to a set which is compact in an appropriate topology. Usually, this is accomplished by the adding of a constraint, as was done in (4.20). Instead, Tikhonov introduced the constraint as a **penalization term** into the functional itself. For example, rather than asking for a solution of the optimization problem

$$\text{Minimize} \quad \|K\psi - f\| \quad \text{subject to} \quad \|\psi\| \leq M, \tag{4.26}$$

we introduce a penalization parameter $\alpha > 0$, and ask for the minimum of the *unconstrained* problem

$$\text{Minimize} \quad \mathcal{J}(\psi) := \|K\psi - f\|^2 + \alpha\,\|\psi\|^2, \quad \psi \in X. \tag{4.27}$$

We expect that the solution of this optimization problem converges to the solution of the equation $K\psi = f$ as α tends to zero. Some of the main properties of Tikhonov's method are collected in the following theorem.

Theorem 4.10. *Let $K : X \longrightarrow Y$ be bounded. For every $\alpha > 0$ there exists a unique minimum ψ^α of $\mathcal{J}$ defined by (4.27) on X. Furthermore, ψ^α satisfies the* **normal equation**

$$\alpha\,(\psi^\alpha, \varphi)_X + (K\psi^\alpha - f, K\varphi)_Y = 0 \quad \textit{for all } \varphi \in X, \tag{4.28a}$$

or, using the adjoint $K^* : Y \longrightarrow X$ of K,

$$\alpha\,\psi^\alpha + K^*K\psi^\alpha = K^*f. \tag{4.28b}$$

If, in addition, K is one-to-one and $\psi \in X$ is the (unique) solution of the equation $K\psi = f$ then $\psi^\alpha \to \psi$ as α tends to zero.

Finally, if $\psi \in K^*(Y)$ or $\psi \in K^*K(X)$, then there exists $c > 0$ with

$$\|\psi^\alpha - \psi\| \leq c\sqrt{\alpha} \quad \textit{or} \quad \|\psi^\alpha - \psi\| \leq c\,\alpha, \quad \textit{respectively.} \tag{4.29}$$

Proof: Let $\{\psi_n\}_{n=1}^\infty \subset X$ be a minimizing sequence, i.e. $\mathcal{J}(\psi_n) \to J := \inf\{\mathcal{J}(\varphi) : \varphi \in X\}$. We show that it is a Cauchy sequence and, therefore, converges. Application of the binomial formula yields

$$\begin{aligned} \mathcal{J}(\psi_n) + \mathcal{J}(\psi_m) &= 2\,\mathcal{J}\left(\frac{1}{2}(\psi_n + \psi_m)\right) + \frac{1}{2}\,\|K(\psi_n - \psi_m)\|^2 \\ &\quad + \frac{\alpha}{2}\,\|\psi_n - \psi_m\|^2 \\ &\geq 2J + \frac{\alpha}{2}\,\|\psi_n - \psi_m\|^2 . \end{aligned}$$

The left hand side converges to $2J$ as n, m tend to infinity. This shows that $\{\psi_n\}_1^\infty$ is a Cauchy sequence and thus convergent. Let $\psi^\alpha = \lim_{n\to\infty} \psi_n$. From the continuity of $\mathcal{J}$ we conclude that $\mathcal{J}(\psi_n) \to \mathcal{J}(\psi^\alpha)$, i.e. $\mathcal{J}(\psi^\alpha) = J$. This proves the existence of a minimum of $\mathcal{J}$.

Now we use the following formula (see proof of Theorem 4.6):

$$\begin{aligned}\mathcal{J}(\psi) - \mathcal{J}(\psi^\alpha) &= 2\,\mathrm{Re}\left(K\psi^\alpha - f, K(\psi - \psi^\alpha)\right) \;+\; 2\alpha\,\mathrm{Re}\,(\psi^\alpha, \psi - \psi^\alpha) \\ &\quad + \;\|K(\psi - \psi^\alpha)\|^2 \;+\; \alpha\,\|\psi - \psi^\alpha\|^2 \\ &= 2\,\mathrm{Re}\left(K^*(K\psi^\alpha - f) + \alpha\psi^\alpha, \psi - \psi^\alpha\right) \\ &\quad + \;\|K(\psi - \psi^\alpha)\|^2 \;+\; \alpha\,\|\psi - \psi^\alpha\|^2\end{aligned}$$

for all $\psi \in X$. From this, the equivalence of the normal equation (4.28b) with the minimization problem for $\mathcal{J}$ is shown exactly as in the proof of Theorem 4.6. Finally, we show that $\alpha I + K^*K$ is one-to-one for every $\alpha > 0$. Let $\alpha\psi + K^*K\psi = 0$. Multiplication by ψ yields $\alpha(\psi,\psi) + (K\psi, K\psi) = 0$, i.e. $\psi = 0$. This proves the first part of the theorem.

Now we study the convergence properties of the solutions ψ^α as α tends to zero. Assume that K is one-to-one and $K\psi = f$. From (4.28b) we see that

$$\alpha(\psi^\alpha - \psi) \;+\; K^*K(\psi^\alpha - \psi) \;=\; -\alpha\,\psi$$

and thus by multiplication by $\psi^\alpha - \psi$:

$$\alpha\,\|\psi^\alpha - \psi\|^2 \;+\; \|K(\psi^\alpha - \psi)\|^2 \;=\; -\alpha\,(\psi, \psi^\alpha - \psi)\,. \tag{4.30}$$

Let us first consider the case where $\psi \in K^*(Y)$, i.e, $\psi = K^*z$ for some $z \in Y$. Then (4.30) becomes

$$\begin{aligned}\alpha\,\|\psi^\alpha - \psi\|^2 \;+\; \|K(\psi^\alpha - \psi)\|^2 &= -\alpha\big(z, K(\psi^\alpha - \psi)\big) \\ &\le \alpha\,\|z\|\;\|K(\psi^\alpha - \psi)\|\,.\end{aligned} \tag{4.31}$$

From this we conclude, first, that $\|K(\psi^\alpha - \psi)\| \le \alpha\,\|z\|$ and, second, that

$$\|\psi^\alpha - \psi\|^2 \;\le\; \|z\|\;\|K(\psi^\alpha - \psi)\| \;\le\; \alpha\,\|z\|^2\,,$$

i.e. $\|\psi^\alpha - \psi\| \le \sqrt{\alpha}\,\|z\|$ which proves the first estimate in (4.29).

Now let $\psi = K^*Kv$ for some $v \in X$. We set $\varphi := Kv$. From (4.28b) we see that $\psi^\alpha = K^*\varphi^\alpha$ where $\varphi^\alpha := \frac{1}{\alpha}K(\psi - \psi^\alpha)$. In terms of φ and φ^α the normal equation (4.28b) takes the form

$$\alpha\,K^*\varphi^\alpha \;+\; K^*K\,K^*(\varphi^\alpha - \varphi) \;=\; 0\,,$$

i.e.

$$\alpha\,\varphi^\alpha \;+\; K\,K^*(\varphi^\alpha - \varphi) \;=\; \rho_\alpha$$

for some $\rho_\alpha \in \mathcal{N}(K^*)$. As above, we subtract $\alpha\varphi$ on both sides and multiply with $\varphi^\alpha - \varphi$. This yields

$$\begin{aligned}\alpha\,\|\varphi^\alpha - \varphi\|^2 \;+\; \|K^*(\varphi^\alpha - \varphi)\|^2 &= \left(\rho_\alpha, \varphi^\alpha - \varphi\right) \;-\; \alpha\,(\varphi, \varphi^\alpha - \varphi) \\ &= \left(\rho_\alpha, \varphi^\alpha - \varphi\right) \;-\; \alpha\left(v, K^*(\varphi^\alpha - \varphi)\right).\end{aligned}$$

The term $(\rho_\alpha, \varphi^\alpha - \varphi)$ vanishes since $\rho_\alpha \in \mathcal{N}(K^*) = \mathcal{R}(K)^\perp$ and $\varphi^\alpha - \varphi \in \mathcal{R}(K)$. The Cauchy-Schwarz inequality yields

$$\|\psi^\alpha - \psi\| \;=\; \|K^*(\varphi^\alpha - \varphi)\| \;\leq\; \alpha\,\|v\|$$

which proves the second estimate of (4.29).

Convergence $\psi^\alpha \to \psi$ without any assumed regularity on ψ is proven by a density argument. Indeed, we have just shown that

$$(\alpha I + K^*K)^{-1} K^*K\psi \longrightarrow \psi \quad \text{for all } \psi \in K^*(Y)\,,$$

i.e. pointwise convergence of the operators on the dense set $K^*(Y)$. Furthermore, from the optimality of ψ^α we conclude that

$$\alpha\,\|\psi^\alpha\|^2 \;\leq\; \mathcal{J}(\psi^\alpha) \;\leq\; \mathcal{J}(\psi) \;=\; \alpha\,\|\psi\|^2\,,$$

i.e., $\|\psi^\alpha\| \;\leq\; \|\psi\|$ for every $\alpha > 0$ and $\psi \in X$. This shows that the operators $(\alpha I + K^*K)^{-1}K^*K$ are uniformly bounded with respect to α. Therefore, we have shown that: (i) there exists a constant $c \geq 1$ with $\left\|(\alpha I + K^*K)^{-1}K^*K\right\| \leq c$ for all α and, (ii), $(\alpha I + K^*K)^{-1}K^*K\psi \longrightarrow \psi$ for all $\psi \in K^*(Y)$. This implies convergence for all $\psi \in X$.[3] Indeed, let $\psi \in X$ and $\varepsilon > 0$ be arbitrary. Choose first $\hat\psi \in K^*(Y)$ with $\left\|\psi - \hat\psi\right\|_X \leq \varepsilon/(3c)$ and then $\alpha_0 > 0$ such that $\left\|\alpha I + K^*K)^{-1}K^*K\hat\psi - \hat\psi\right\|_X \leq \varepsilon/3$ for all $\alpha \leq \alpha_0$. Then, by the triangle inequality,

$$\begin{aligned}
\left\|(\alpha I + K^*K)^{-1}K^*K\psi - \psi\right\|_X &\leq \|(\alpha I + K^*K)^{-1}K^*K(\psi - \hat\psi)\|_X \\
&\quad + \|(\alpha I + K^*K)^{-1}K^*K\hat\psi - \hat\psi\|_X \\
&\quad + \left\|\hat\psi - \psi\right\|_X \\
&\leq \varepsilon/3 \;+\; \varepsilon/3 \;+\; \varepsilon/(3c) \;\leq\; \varepsilon \quad \text{for all } \alpha \leq \alpha_0\,.
\end{aligned}$$

This estimate implies convergence for every $\psi \in X$. □

The assumptions $\psi \in K^*(Y)$ or $\psi \in K^*K(X)$ are *smoothness assumptions* on the solution ψ since in concrete examples (see, e.g., Example 4.11) the operator K^* and K^*K are smoothing in the sense that the ranges of both of these operators contain only functions of greater regularity then those in X.

We will illustrate this method by an example, originally investigated by Tikhonov [135].

Example 4.11. We consider the following integral equation of the first kind:

[3] This is actually one part of the Theorem of Banach-Steinhaus (see [145]).

$$\int_0^1 g(t,s)\,\psi(s)\,ds \;=\; f(t)\,, \quad t \in [a,b]\,, \tag{4.32}$$

where $g : [a,b] \times [0,1] \longrightarrow \mathbb{R}$ and $f : [a,b] \longrightarrow \mathbb{R}$ are continuous functions. The operator K is therefore defined by

$$(K\psi)(t) \;:=\; \int_0^1 g(t,s)\,\psi(s)\,ds \quad t \in [a,b]\,.$$

For X and Y we take the **Sobolev space** $X = H^1(0,1)$ and $Y := L^2(a,b)$, respectively (see Definition A.20). The operator K, considered as an operator from $L^2(0,1)$ into $L^2(a,b)$ is compact by Theorem A.38. Since $H^1(0,1)$ is boundedly (even compactly) imbedded in $L^2(0,1)$ the operator K is also compact considered as an operator from $H^1(0,1)$ into $L^2(a,b)$. The Tikhonov functional $\mathcal{J}$ and the normal equation for the minimum $\psi^\alpha \in H^1(0,1)$ of $\mathcal{J}$ take the form

$$\mathcal{J}(\psi) \;=\; \|K\psi - f\|^2_{L^2(a,b)} \;+\; \alpha \int_0^1 \left[\psi(t)^2 \;+\; \psi'(t)^2\right] dt\,, \quad \psi \in H^1(0,1)\,,$$

$$\left(K\psi^\alpha - f, K\varphi\right)_{L^2(a,b)} \;+\; \alpha\left(\psi^\alpha, \varphi\right)_X \;=\; 0 \quad \text{for all } \varphi \in H^1(0,1)\,,$$

respectively. We rewrite the last equation as

$$\left(K^*(K\psi^\alpha - f), \varphi\right)_{L^2(0,1)} \;+\; \alpha\left(\psi^\alpha, \varphi\right)_{L^2(0,1)} \;+\; \alpha\left((\psi^\alpha)', \varphi'\right)_{L^2(0,1)} \;=\; 0$$

for all $\varphi \in H^1(0,1)$, where $K^* : L^2(a,b) \longrightarrow L^2(0,1)$ is the L^2–adjoint of K, i.e.

$$\left(K^*\psi\right)(t) \;:=\; \int_a^b g(s,t)\,\psi(s)\,ds\,, \quad t \in [0,1]\,.$$

From the **Fundamental Lemma of the Calculus of Variations** (Lemma 4.12 below) we conclude that even $\psi^\alpha \in H^2(0,1)$. Furthermore, it follows from the Fundamental Lemma that $\frac{d}{dt}\psi^\alpha(0) = \frac{d}{dt}\psi^\alpha(1) = 0$ and

$$K^*\left(K\psi^\alpha - f\right) \;+\; \alpha\,\psi^\alpha \;-\; \alpha\,\frac{d^2}{dt^2}\psi^\alpha \;=\; 0\,,$$

i.e. $\psi^\alpha \in H^2(0,1)$ solves an **integro-differential equation**. Note that, from the smoothness of the kernel g, the term $K^*(K\psi^\alpha - f)$ is continuous so that the solution ψ^α is even twice continuously differentiable on $[0,1]$.

Lemma 4.12. *Let $g, h \in L^2(0,1)$ and*

$$(g,\varphi)_{L^2(0,1)} \;+\; \left(h, \varphi'\right)_{L^2(0,1)} \;=\; 0 \quad \textit{for all } \varphi \in H^1(0,1)\,. \tag{4.33}$$

Then $h \in H^1(0,1)$, $h(0) = h(1) = 0$, and $g = h'$.

Proof: Define $\tilde{h}(t) := \int_0^t g(s)\,ds$, $t \in [0,1]$. Then $\tilde{h} \in H^1(0,1)$ and $\tilde{h}(0) = 0$ and $\tilde{h}' = g$. We will show that $\tilde{h} = h$ and $h(1) = 0$.

Substitution of $g = \tilde{h}'$ in (4.33) and integration by parts of the first term yields

$$\begin{aligned} 0 &= \left(\tilde{h}', \varphi\right)_{L^2(0,1)} \;+\; (h, \varphi')_{L^2(0,1)} \\ &= \tilde{h}(1)\varphi(1) \;-\; \left(\tilde{h}, \varphi'\right)_{L^2(0,1)} \;+\; (h, \varphi')_{L^2(0,1)}\,, \end{aligned} \tag{4.34}$$

for all $\varphi \in H^1(0,1)$. Now we substitute $\varphi(t) := -\int_t^1 [h(s) - \tilde{h}(s)]\,ds$ into this equation. From $\varphi(1) = 0$ and $\varphi' = h - \tilde{h}$ we conclude that $\|h - \tilde{h}\|^2_{L^2(0,1)} = 0$, i.e. $h = \tilde{h}$. Since (4.34) holds for all $\varphi \in H^1(0,1)$ we see, finally, that $h(1) = 0$. □

The previous theorem assumes that the right hand side f is known *exactly*, i.e. without noise. In this unrealistic case, α should be taken as small as possible, ideally $\alpha = 0$. In general, however, one knows only an approximation $f^\delta \in Y$ of f and the noise level δ with $\|f^\delta - f\| \le \delta$. Therefore, instead of computing ψ^α from (4.28) we compute $\psi^{\delta,\alpha}$ from

$$\alpha\,\psi^{\delta,\alpha} \;+\; K^*K\psi^{\delta,\alpha} \;=\; K^*f^\delta\,. \tag{4.35}$$

The error is decomposed and then estimated by

$$\|\psi^{\delta,\alpha} - \psi\| \;\le\; \|\psi^{\delta,\alpha} - \psi^\alpha\| \;+\; \|\psi^\alpha - \psi\|\,. \tag{4.36}$$

The first term compares solutions for equation (4.35) with right hand sides K^*f^δ and K^*f, while the second term is independent of the noise δ and has been studied in the previous theorem.

In order to estimate the first term we have to compute a bound on the norm $\|(\alpha I + K^*K)^{-1}K^*\|$. If $\varphi \in X$ is the solution of

$$\alpha\,\varphi \;+\; K^*K\varphi \;=\; K^*h \quad \text{for some } h \in Y$$

then, since φ is the minimum of $\mathcal{J}(\psi) = \alpha\,\|\psi\|^2 + \|K\psi - h\|^2$ on X,

$$\alpha\,\|\varphi\|^2 \;\le\; \alpha\,\|\varphi\|^2 \;+\; \|K\varphi - h\|^2 \;\le\; \alpha\,\|0\|^2 \;+\; \|K0 - h\|^2 \;=\; \|h\|^2\,.$$

This proves $\|\varphi\| \le \|h\|/\sqrt{\alpha}$ and thus

$$\|(\alpha I + K^*K)^{-1}K^*\| \;\le\; \frac{1}{\sqrt{\alpha}}\,.$$

We have derived the following basic estimate

$$\begin{aligned}\|\psi^{\delta,\alpha} - \psi\| &\le \|(\alpha I + K^*K)^{-1}K^*(f^\delta - f)\| \; + \; \|\psi^\alpha - \psi\| \\ &\le \frac{\delta}{\sqrt{\alpha}} \; + \; \|\psi^\alpha - \psi\| \,. \end{aligned} \tag{4.37}$$

In the case $\psi \in K^*(Y)$ we conclude that

$$\|\psi^{\delta,\alpha} - \psi\| \;\le\; \frac{\delta}{\sqrt{\alpha}} \;+\; c\sqrt{\alpha}\,. \tag{4.38}$$

This estimate suggests the choice $\alpha = \alpha(\delta) = c_1\,\delta$ for some $c_1 > 0$. Indeed, $\alpha = \delta/c$ is the minimum of the right hand side of (4.38) considered as a function of α. With this choice of $\alpha(\delta)$ the error is

$$\|\psi^{\delta,\alpha} - \psi\| \;\le\; \tilde{c}\sqrt{\delta} \tag{4.39}$$

for some $\tilde{c} > 0$. In the case $\psi \in K^*K(X)$ we conclude from (4.37) that

$$\|\psi^{\delta,\alpha} - \psi\| \;\le\; \frac{\delta}{\sqrt{\alpha}} \;+\; c\,\alpha \tag{4.40}$$

which leads to $\alpha = \alpha(\delta) = c_2\,\delta^{2/3}$. With this choice of $\alpha(\delta)$ the error is

$$\|\psi^{\delta,\alpha} - \psi\| \;\le\; \tilde{c}\,\delta^{2/3}\,. \tag{4.41}$$

These choices of the regularization parameter α can be made *a priori*, i.e. before starting the computation. It is convenient, however, to determine α *a posteriori*, i.e. during the computational process. A common method is **Morozov's discrepancy principle**, which determines $\alpha = \alpha(\delta) > 0$ such that the corresponding $\psi^{\delta,\alpha}$ of (4.35) satisfies

$$\|K\psi^{\delta,\alpha} - f^\delta\| \;=\; \tau\,\delta \tag{4.42}$$

where $\tau > 1$ is some fixed parameter. For more on this principle we refer to [67] or the original literature [101, 102] and, for modifications, [37, 116].

4.5 The Synthesis Problem for the Finite Linear Line Source

In this section we apply the theoretical results of the previous sections to the synthesis problem for the linear line source. The application of the regularization methods lead, via the normal equations, to the *numerical* problem of solving a Fredholm integral equation of the second kind. It will therefore be useful to devote a subsection (Subsection 4.5.2) to the Nyström method before studying the normal equations. In the final subsection we return to specific examples of synthesis problems which have been discussed in the literature.

4.5.1 Basic Equations

We will consider not only the case of the operator $K : L^2(-\ell,+\ell) \longrightarrow C[-1,+1]$ given by (4.2) but allow the operator to have the more general form

$$(K\psi)(t) := \alpha(t) \int_{-\ell}^{\ell} \psi(s)\, \mathrm{e}^{-ikts}\, ds\,, \quad |t| \leq 1\,, \tag{4.43}$$

where $k = 2\pi/\lambda$ is the wave number, $\lambda > 0$ the wave length, and $\alpha \in C[-1,+1]$ positive on $(-1,+1)$. We think of $\alpha \equiv 1$ or $\alpha(t) = \sqrt{1-t^2}$, $|t| \leq 1$, as the particularly interesting cases.

The synthesis problem leads to the integral equation $K\psi = f$ i.e.,

$$\alpha(t) \int_{-\ell}^{\ell} \psi(s)\, \mathrm{e}^{-ikts}\, ds = f(t)\,, \quad |t| \leq 1\,.$$

All the methods of regularization described in Sections 4.3 and 4.4 lead to equations of the type

$$\eta\psi + K^*K\psi = \rho K^* f\,.$$

A simple change of the order of integration shows that the operators K^* and K^*K are given by

$$\big(K^*g\big)(t) = \int_{-1}^{1} g(s)\, \alpha(s)\, \mathrm{e}^{ikts}\, ds\,, \quad |t| \leq \ell\,,$$

$$\big(K^*K\psi\big)(t) = \int_{-\ell}^{\ell} \psi(s)\, a(t-s)\, ds\,, \quad |t| \leq \ell\,, \quad \text{where}$$

$$a(\tau) = \int_{-1}^{1} \alpha(s)^2\, \mathrm{e}^{iks\tau}\, ds\,, \quad \tau \in \mathbb{R}\,.$$

For the special cases $\alpha \equiv 1$ and $\alpha(s) = \sqrt{1-s^2}$ the kernel a takes the forms

$$a(\tau) = 2\mathrm{sinc}\,(k\tau) \quad \text{and} \quad a(\tau) = \frac{4}{(k\tau)^3}\big[\sin(k\tau) - (k\tau)\cos(k\tau)\big]\,, \text{ respectively.}$$

The function a is always analytic on $\mathbb{R}$.

We are thus led to the problem of solving **Fredholm integral equations** of the second kind. Only in very special cases such integral equations can be solved explicitly but in most cases one is forced to apply numerical methods for their solution. We turn to this problem in the next subsection of this part.

4.5.2 The Nyström Method

In this subsection we study Fredholm integral equations of the second kind in the form

$$\psi(t) \;+\; \int_a^b g(t,s)\,\psi(s)\,ds \;=\; h(t)\,, \quad t \in [a,b]\,, \tag{4.44}$$

where $g : [a,b] \times [a,b] \longrightarrow \mathbb{C}$ and $h : [a,b] \longrightarrow \mathbb{C}$ are continuous functions. In the Nyström method one replaces the integral by some quadrature rule of the kind which we discussed in Subsection 1.5.3:

$$I(\varphi) \;:=\; \int_a^b \varphi(s)\,ds \;\approx\; Q_n(\varphi) \;:=\; \sum_{j=1}^n w_j^{(n)}\,\varphi\bigl(t_j^{(n)}\bigr) \tag{4.45}$$

for given nodes $t_j = t_j^{(n)} \in [a,b]$ and weights $w_j = w_j^{(n)} \in \mathbb{R}$, $j = 1,\dots,n$. The Nyström method replaces the integral equation (4.44) by

$$\psi^{(n)}(t) \;+\; \sum_{j=1}^n w_j\, g(t,t_j)\,\psi^{(n)}(t_j) \;=\; h(t)\,, \quad t \in [a,b]\,. \tag{4.46}$$

Note that this is still an equation for the *function* $\psi^{(n)}$. However, it is equivalent to the finite linear system

$$\psi^{(n)}(t_p) \;+\; \sum_{j=1}^n w_j\, g(t_p,t_j)\,\psi^{(n)}(t_j) \;=\; h(t_p)\,, \quad p = 1,\dots,n\,, \tag{4.47}$$

in the sense that with any solution $(\psi_j)_{j=1}^n \in \mathbb{C}^n$ of (4.47) the **Nyström interpolant**

$$\psi^{(n)}(t) \;:=\; h(t) \;-\; \sum_{j=1}^n w_j\, g(t,t_j)\,\psi_j\,, \quad t \in [a,b]\,, \tag{4.48}$$

defines a solution of (4.46). The convergence analysis for the Nyström method can be found in, e.g., [74]. The error $\|\psi^{(n)} - \psi\|_{C[a,b]}$ depends on the smoothness of the solution ψ and the quadrature rule. In particular, the following result is well known.

Theorem 4.13. *Assume the quadrature formulae (4.45) to be convergent for every continuous function φ. Assume, furthermore, that the integral equation (4.44) is uniquely solvable for every $h \in C[a,b]$. Then, for sufficiently large n, the equations (4.46) are uniquely solvable in $C[a,b]$. For the solutions ψ of (4.44) and $\psi^{(n)}$ of (4.46) we have the error estimate*

$$\|\psi^{(n)} - \psi\|_{C[a,b]} \;\le\; c \max_{a\le t\le b}\bigl|(Q_n - I)\bigl(g(t,\cdot)\,\psi\bigr)\bigr|\,, \quad n \ge n_0\,, \tag{4.49}$$

for some $c > 0$.

We illustrate this result by considering three examples of quadrature formulae which correspond to the examples of Subsection 1.5.3.

First, assume that $g : \mathbb{R} \times \mathbb{R} \longrightarrow \mathbb{C}$ and $h : \mathbb{R} \longrightarrow \mathbb{C}$ are analytic and $2\pi-$periodic. For g we assume this with respect to both variables. We study the integral equation

$$\psi(t) + \int\limits_0^{2\pi} g(t,s)\, \psi(s)\, ds = h(t)\,, \quad t \in [0, 2\pi]\,. \tag{4.50}$$

As indicated in Subsection 1.5.3, we should take the trapezoidal rule, i.e. $n = 2N$, $w_j = \pi/N$, and $t_j = j\pi/N$ for $j = 0, \ldots, 2N-1$. Assume again, that the integral equation (4.50) has a unique periodic solution $\psi \in C[0, 2\pi]$ for every periodic function $h \in C[0, 2\pi]$. The integral equation (4.50) yields immediately that for analytic functions h, the solution ψ is necessarily also analytic. Therefore, for analytic functions h, the application of Theorems 1.17 and 4.13 yields exponential order of convergence:

$$\|\psi^{(N)} - \psi\|_{C[0,2\pi]} \leq \frac{c}{\mathrm{e}^{\sigma N} - 1}\,, \quad N \geq N_0\,, \tag{4.51}$$

for some $c > 0$ and $\sigma > 0$.

The assumption of periodicity is satisfied for *closed loop antennas*. For *open line sources*, the functions g and h fail to be periodic. In this case, the Gauss-Legendre method

$$\int\limits_{-1}^{1} \varphi(t)\, dt \approx \sum_{j=1}^{n} w_j\, \varphi(t_j) \tag{4.52}$$

is better suitable where we take t_j as the zeros of the Legendre polynomial P_n which are all simple and lie in the interval $(-1, 1)$. The weights w_j are known to be strictly positive. Then an analogous error estimate (4.51) holds for analytic data g and h.

As a third example of a quadrature rule we had applied the trapezoidal rule to the transformed integral

$$\int\limits_a^b \varphi(s)\, ds = \int\limits_0^{2\pi} \tilde{\varphi}(t)\, dt \quad \text{with} \quad \tilde{\varphi}(t) = w'(t)\, \varphi\big(w(t)\big)\,, \; 0 \leq t \leq 2\pi\,. \tag{4.53}$$

Here, the function $w : [0, 2\pi] \longrightarrow [a, b]$ is chosen to be bijective, strictly monotonically increasing and infinitely often differentiable with derivatives

$$w^{(j)}(0) = w^{(j)}(2\pi) = 0\,, \quad j = 1, \ldots, p-1\,,$$

for some odd $p \geq 3$. Application of (1.83) and Theorem 4.13 then yields the order $p - 2$ of convergence:

$$\|\psi^{(N)} - \psi\|_{C[a,b]} \leq \frac{c}{N^{p-2}}, \quad N \geq N_0, \tag{4.54}$$

for some $c > 0$ provided g and h are smooth enough.

In the following subsection we will apply the Nyström method to the integral equations which arose in Sections 4.3 and 4.4. We note that all of the theoretical results remain valid if the right hand side h in (4.46) is replaced by a function h_n provided $\|h_n - h\|_{C[a,b]} \to 0$ as $n \to \infty$. The error estimates (4.49), (4.51), and (4.54) have to be replaced by

$$\|\psi^{(n)} - \psi\|_{C[a,b]} \leq c \left[\max_{a \leq t \leq b} \left|(Q_n - I)\big(g(t,\cdot)\,\psi\big)\right| \;+\; \|h_n - h\|_{C[a,b]} \right] \tag{4.55a}$$

$$\|\psi^{(N)} - \psi\|_{C[0,2\pi]} \leq \frac{c}{e^{\sigma N} - 1} \quad \text{if } \|h_N - h\|_{C[0,2\pi]} \leq \frac{c'}{e^{\sigma N} - 1}, \tag{4.55b}$$

$$\|\psi^{(N)} - \psi\|_{C[a,b]} \leq \frac{c}{N^{p-2}} \quad \text{if } \|h_N - h\|_{C[a,b]} \leq \frac{c'}{N^{p-2}}. \tag{4.55c}$$

4.5.3 Numerical Solution of the Normal Equations

First, we observe that the normal equations (4.21), (4.25), and (4.28b), for the restricted quasi-solution, the restriction on the super-gain ratio, and the Tikhonov regularization all take the form $\eta\,\psi + K^*K\psi = \rho\,K^*f$, i.e.

$$\eta\,\psi(t) \;+\; \int_{-\ell}^{\ell} \psi(s)\,a(t-s)\,ds \;=\; \rho \int_{-1}^{1} f(s)\,e^{ikst}\,ds\,, \quad |t| \leq \ell, \tag{4.56}$$

with different meanings of $\eta \neq 0$ and ρ. With $g(t,s) := a(t-s)$ and $h(t) := \int_{-1}^{1} f(s)\,\alpha(s)\,e^{ikst}\,ds$ the equation takes the form (4.44) (after an obvious division by η). For the Nyström method we use the Gauss-Legendre rule, i.e.

$$\int_{-\ell}^{\ell} \varphi(s)\,ds \;=\; \ell \int_{-1}^{1} \varphi(\tau\ell)\,d\tau \;\approx\; \ell \sum_{j=1}^{n} w_j\,\varphi(t_j\ell)$$

where w_j, $t_j \in (-1,+1)$ are the Gauss-Legendre weights for the interval $(-1,+1)$.

In the cases where $h(t)$ can not be computed analytically we approximate the integral also by the Gauss-Legendre rule, i.e.

$$h(t) \;:=\; \int_{-1}^{1} f(s)\,\alpha(s)\,e^{ikst}\,ds \;\approx\; h_n(t) \;:=\; \sum_{j=1}^{n} w_j\,f(t_j)\,\alpha(t_j)\,e^{ikt_jt}\,.$$

(In the case $\alpha(t) = \sqrt{1-t^2}$ one should the Gauss-Tschebycheff rule instead which takes α as a weight function.) The Nyström equations (4.46) and (4.47) take the form

$$\eta\,\psi^{(n)}(t\ell) \;+\; \ell\sum_{j=1}^{n} w_j\,\psi^{(n)}(t_j\ell)\,a\big[\ell(t-t_j)\big] = \rho\,h_n(t\ell)\,, \qquad |t|\leq 1\,, \tag{4.57a}$$

$$\eta\,\psi^{(n)}(t_p\ell) \;+\; \ell\sum_{j=1}^{n} w_j\,\psi^{(n)}(t_j\ell).a\big[\ell(t_p-t_j)\big] = \rho\,h_n(t_p\ell)\,, \tag{4.57b}$$

for $p=1,\ldots,n$, respectively. In the case where $h(t)$ can be computed analytically, we replace h_n by h in equation (4.57b).

In the form (4.57b) the corresponding matrix is not symmetric although the integral equation (4.56) has a symmetric kernel. Therefore, we set $\psi_j := \sqrt{nw_j}\,\psi^{(n)}(t_j\ell)$, $j=1,\ldots,n$, and multiply equation (4.57b) by $\sqrt{nw_p}$. This yields

$$\eta\,\tilde{\psi} \;+\; A\tilde{\psi} \;=\; r\,, \tag{4.58}$$

for the vector $\tilde{\psi} = (\psi_j)_{j=1}^n \in \mathbb{C}^n$ with the symmetric matrix A and vector r whose entries are given by

$$A_{p,j} \;:=\; \ell\,a\big[\ell(t_p-t_j)\big]\,\sqrt{w_p\,w_j} \quad\text{and}\quad r_p \;:=\; \rho\,\sqrt{nw_p}\,h_n(t_p\ell)\,. \tag{4.59}$$

The norm $\|\psi\|^2_{L^2(-\ell,+\ell)}$ and the defect $\|K\psi-f\|^2_{L^2(-1,+1)}$ are replaced by the discrete forms

$$\|\psi\|^2_{L^2(-\ell,+\ell)} \approx \ell\sum_{j=1}^{n} w_j\,\big|\psi^{(n)}(t_j)\big|^2 \;=\; \frac{\ell}{n}\sum_{j=1}^{n}|\psi_j|^2\,,$$

$$\begin{aligned}\|K\psi-f\|^2_{L^2(-1,+1)} &= \|K\psi\|^2_{L^2(-1,+1)} \;-\; 2\,\mathrm{Re}\,(K\psi,f)_{L^2(-1,+1)} \;+\|f\|^2_{L^2(-1,+1)}\\ &= (\rho-2)\,(h,\psi)_{L^2(-\ell,+\ell)} \;-\; \eta\,\|\psi\|^2_{L^2(-\ell,+\ell)} \;+\|f\|^2_{L^2(-1,+1)}\\ &\approx \frac{\ell(\rho-2)}{n\rho}\sum_{j=1}^{n}\psi_j\,\overline{r_j} \;-\; \frac{\eta\ell}{n}\sum_{j=1}^{n}|\psi_j|^2 \;+\; \|f\|^2_{L^2(-1,+1)}\,,\end{aligned}$$

where we have multiplied the normal equation $\eta\,\psi + K^*K\psi = \rho\,K^*f = \rho\,h$ by ψ and substituted $\|\psi\|^2_{L^2(-1,+1)}$.

4.5.4 Applications of the Regularization Techniques

As a first example we implement the method of restricted quasi-solutions (4.20), i.e. we have to solve the normal equation (4.21b)

$$\eta\,\psi^o \;+\; K^*K\psi^o \;=\; K^*f\,, \tag{4.60}$$

and have to determine $\eta>0$ such that $\|\psi^o\|_{L^2(-\ell,+\ell)} = M$.

We consider two special examples. In the first one, the equation $K\psi = f$ is solvable in contrast to the second example where it is not. In both cases the operator K is given by (4.43) for $\alpha\equiv 1$ i.e.,

$$(K\psi)(t) \;:=\; \int_{-\ell}^{\ell} \psi(s)\, \mathrm{e}^{-ikts}\, ds\,, \quad |t| \le 1\,.$$

At the end we show for one of these examples that the results exhibit no significant numerical change if we replace $\alpha \equiv 1$ by $\alpha(t) = \sqrt{1-t^2}$.

In the first example we take $f_1(t) = \big(K\psi_1\big)(t)$, $|t| \le 1$, with $\psi_1(s) = \sin(\pi s/\ell)$, $|s| \le \ell$, i.e.

$$f_1(t) \;=\; (K\psi_1)(t) \;=\; i\ell\left(\operatorname{sinc}\big[\pi + k\ell t)\big] \;-\; \operatorname{sinc}\big[\pi - k\ell t)\big]\right), \quad |t| \le 1\,.$$

For the computation of $h_1(t) = (K^* f_1)(t)$ we use the Gauss-Legendre quadrature rule.

In the second example we take the function considered by Rhodes [114]. Let the original function g given by

$$g(\theta) \;:=\; \begin{cases} 1/\cos\theta\,, \; \theta_1 \le \theta \le \theta_2\,, \\ \quad 0\,, \quad \text{otherwise}, \end{cases} \quad 0 \le \theta \le \pi\,,$$

with $\theta_1 = 15\pi/180$ and $\theta_2 = 75\pi/180$. This transforms into

$$f_2(t) \;=\; \begin{cases} 1/t\,, \; t_1 \le t \le t_2\,, \\ \;0\,, \; \text{otherwise}, \end{cases} \quad |t| \le 1\,,$$

with $t_1 = \cos\theta_2$ and $t_2 = \cos\theta_1$. Here, $h_2 = K^* f_2$ is given by

$$h_2(t) \;=\; (K^* f_2)(t) \;=\; \int_{t_1}^{t_2} \frac{1}{s}\, \mathrm{e}^{ikst}\, ds\,, \quad |t| \le \ell\,,$$

which has also to be computed numerically by the Gauss-Legendre rule (4.52) after transforming the interval $\big(t_1, t_2\big)$ onto $(-1, +1)$.

Figure 4.1 shows the graphs of the discrete versions of $\eta \mapsto \|\psi\|_{L^2(-\ell,\ell)}$ and $\eta \mapsto \|K\psi - f_1\|_{L^2(-1,+1)}$ for $\ell = 1$ and wave length $\lambda = 2$ for the first example. Here we denote be ψ the solution of (4.60) where added 10% random noise on f_1.

Then we determine η with $\|\psi\|_{L^2(-1,+1)} = M$ by the Regula Falsi. For $M = 0.5$ and $M = 1.5$ we computed $\eta \approx 1.27$ and $\eta \approx 2.34 * 10^{-5}$, respectively. The corresponding plots of f_1 and $K\psi^o$ and optimal currents are shown in Figures 4.2 and 4.3, respectively. We note that the "true" current $\psi^o(t) = \sin(\pi s/\ell)$ has norm 1. This explains why the plots for f_1 and $K\psi^o$ in the left of Figure 4.3 are indistinguishable. The Figures 4.4 – 4.6 correspond to Figures 4.1 – 4.3 but this time as a second example, $\delta = 0$, and $M = 1$ and $M = 3$. The corresponding values of the Lagrange multipliers are $\eta \approx 0.25$ and $\eta \approx 0.0057$, respectively.

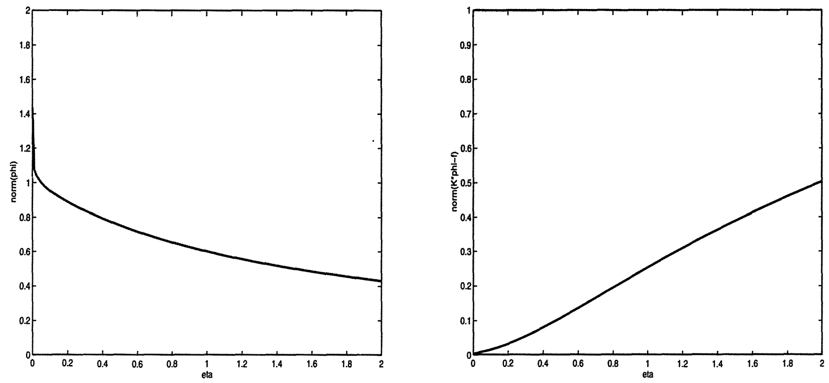


Fig. 4.1. $\eta \mapsto \|\psi^o\|$ and $\eta \mapsto \|K\psi^o - f_1\|$ for the first example and 10% noise

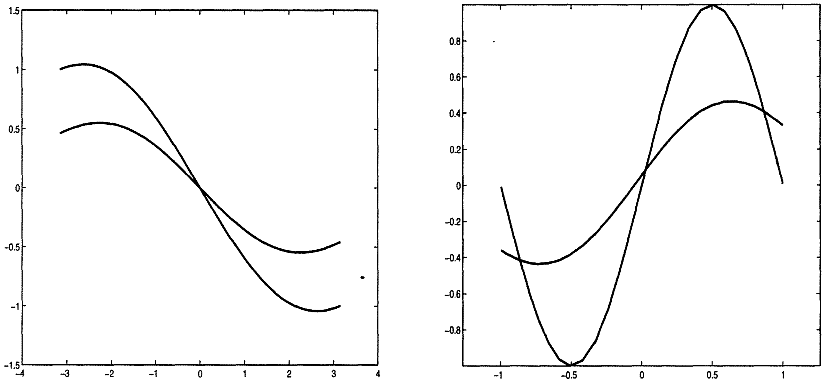


Fig. 4.2. $\operatorname{Im} f_1$, $\operatorname{Im} K\psi^o$ and $\operatorname{Re} \psi^o$, ψ_1 for the first example, 10% noise and $M = 0.5$

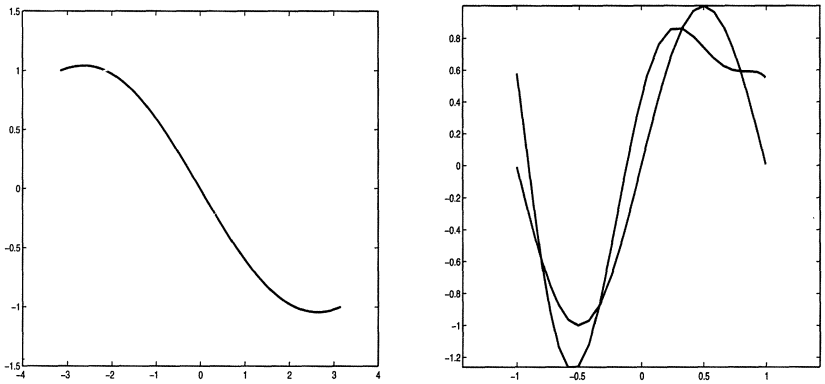


Fig. 4.3. $\operatorname{Im} f_1$, $\operatorname{Im} K\psi^o$ (indistinguishable) and $\operatorname{Re} \psi^o$, ψ_1 for the first example, 10% noise and $M = 1.5$

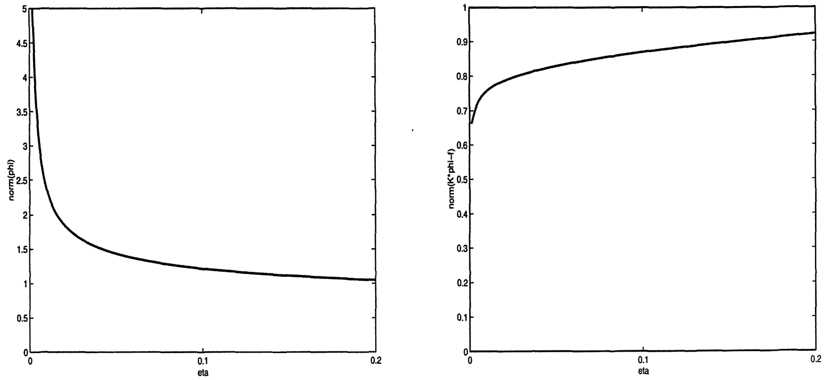


Fig. 4.4. $\eta \mapsto \|\psi\|$ and $\eta \mapsto \|K\psi - g\|$ for the second example (no noise)

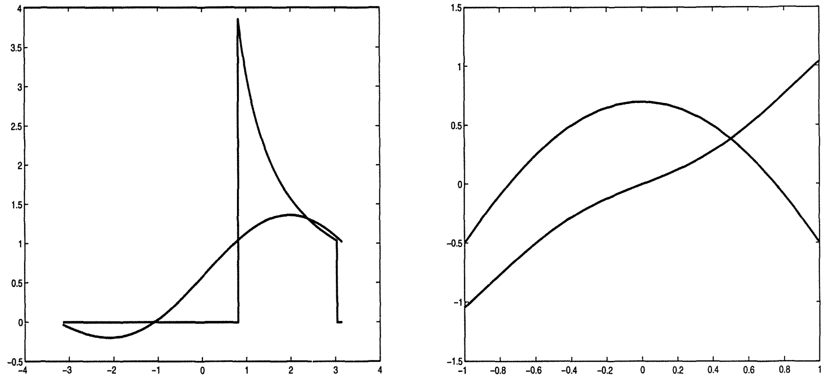


Fig. 4.5. f_2, $\operatorname{Re} K\psi^o$ and $\operatorname{Re}\psi^o$, $\operatorname{Im}\psi^o$ for the second example and $M = 1$, no noise

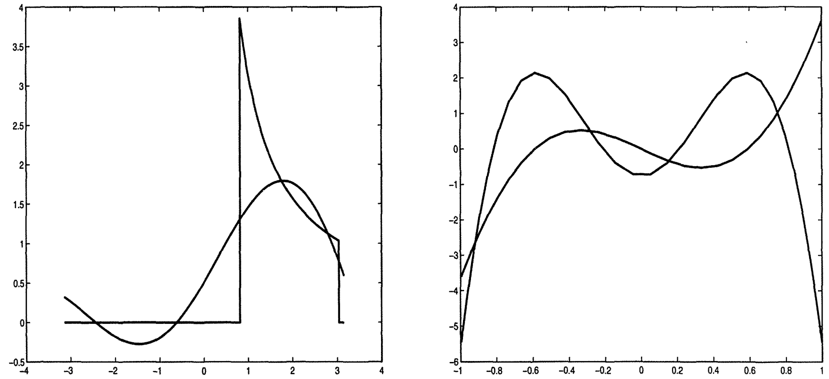


Fig. 4.6. f_2, $\operatorname{Re} K\psi^o$ and $\operatorname{Re}\psi^o$, $\operatorname{Im}\psi^o$ for the second example and $M = 3$, no noise

The second method which we have studied in Section 4.3 is to minimize $\|K\psi - f\|_{L^2(-1,+1)}$ subject to a constraint on the super-gain ratio, i.e. subject to $\|\psi\|_{L^2(-\ell,+\ell)} \leq M \,\|K\psi\|_{L^2(-1,+1)}$ for some $M > 1/\,\|K\|$. We have now to solve equation (4.25) instead of (4.21) and have to determine $\eta > 0$ from the equation $\|\psi\|_{L^2(-\ell,+\ell)} = M\,\|K\psi\|_{L^2(-1,+1)}$. Equation (4.25) differs from (4.21) only in the term involving $K^*K\psi$, i.e. we must replace K^*K in (4.21) and A in (4.59) by $(1 - \eta M^2)K^*K$ and $(1 - \eta M^2)A$, respectively. In Figure 4.7 we show the discrete version of $\eta \mapsto \|\psi\|_{L^2(-\ell,+\ell)}\,/\,\|K\psi\|_{L^2(-1,+1)} \approx |\tilde{\psi}|/\sqrt{\tilde{\psi}^* A\,\tilde{\psi}}$ and $\|K\psi - f_2\|_{L^2(-1,+1)}$ for the second example, wave length $\lambda = 2$, and $M = 1.1$. We determine the Lagrange multiplier as $\eta \approx 0.041$ show plots of f_2, $K\psi$ and the corresponding current in Figure 4.8.

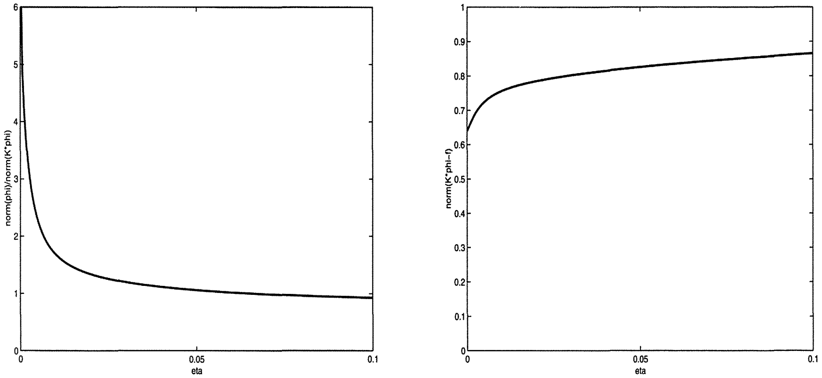


Fig. 4.7. $\eta \mapsto \|\psi\|\,/\,\|K\psi\|$ and $\eta \mapsto \|K\psi - g\|$ for $M = 1.1$

Finally, we may consider this last example also for the case where K is given by

$$(K\psi)(t) \;:=\; \sqrt{1-t^2}\int\limits_{-\ell}^{\ell} \psi(s)\,\mathrm{e}^{-ikts}\,ds\,, \qquad |t| \leq 1\,.$$

Again, $\lambda = 2$ and $M = 1.1$. The result is shown in Figure 4.9.

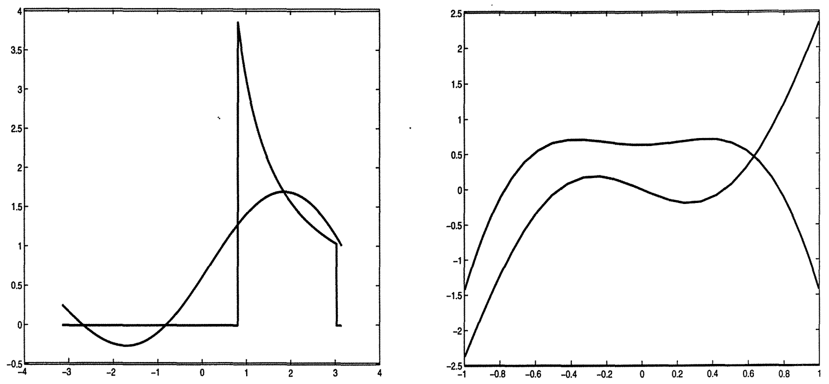


Fig. 4.8. f_2, $\operatorname{Re} K\psi^o$ and $\operatorname{Re}\psi^o$, $\operatorname{Im}\psi^o$ for the second example and $M = 1.1$, no noise

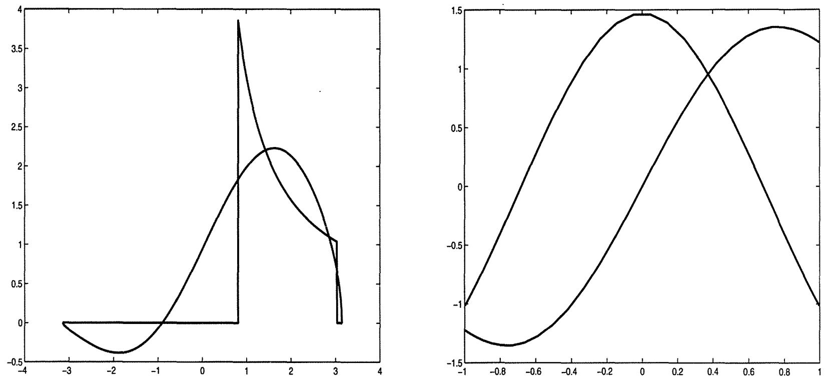


Fig. 4.9. The same as in Figure 4.8 but for the operator K with $\alpha(t) = \sqrt{1-t^2}$

5

Boundary Value Problems for the Two-Dimensional Helmholtz Equation

It is the aim of this chapter to introduce the reader to the *mathematical* theory of boundary value problems. A typical boundary value problem consists of a *differential equation* which the unknown field has to satisfy in some region Ω, certain *boundary conditions* on the boundary $\partial\Omega$ of Ω and a *decay condition* at infinity if Ω is unbounded. In electromagnetic theory we study the Maxwell equations for the time-harmonic case (2.13a)–(2.13d) from Section 2.7 with the boundary conditions specified in Section 2.11, and the radiation condition of Section 2.9. We will treat the corresponding boundary value problems for Maxwell's equations in the next Chapter 6. Here, as an introduction, we will consider the important E- and H-modes introduced in Section 2.8. We do not try to be exhaustive but rather present the basic ideas of the integral equation method.[1] We start by considering smooth boundary data, which allows us to apply the classical Green's theorems. For a rigorous study of boundary integral equation methods for Helmholtz and Maxwell equations and the proofs of the theorems we refer the reader to [29]. In Section 5.4 we will then sketch an approach for L^2-boundary data. These non-smooth boundary data are important for studying questions of existence of optimal solutions of antenna problems.

5.1 Introduction and Formulation of the Problems

In this chapter we will study the following two model problems which correspond to the E- and H-mode for impedance and perfectly conducting

[1] We have seen in Chapter 1 that the analysis of arrays assumes either that the spacing of elements is chosen so that mutual coupling effects are very small or that the question is ignored. It is important to emphasize that the appropriate use of integral equations *automatically* include such effects. Moreover, by the use of boundary integral equation methods the dimension of the space to be discretized for numerical simulations is reduced by one.

boundary conditions (cf. Section 2.11). We restrict ourselves to the physically most important case: that the antenna, described by an infinite cylinder in x_3−direction is imbedded in free space.

Let $\Omega \subset \mathbb{R}^2$ be a bounded region with connected exterior $\Omega^c := \mathbb{R}^2 \setminus \overline{\Omega}$ and smooth boundary $\partial\Omega$. By the latter we mean $\partial\Omega \in C^2$ in the sense defined, e.g., in [29]. We denote by $\boldsymbol{n}(\boldsymbol{x})$ the unit normal vector in $\boldsymbol{x} \in \partial\Omega$ directed into the exterior of Ω. Let $k > 0$ be given. We recall from (2.15) that, in physically relevant situations, k is the free space wave number $k = \omega\sqrt{\mu_0\epsilon_0}$. Furthermore, let g and λ be continuous functions on $\partial\Omega$. We set $D_R := \{\boldsymbol{x} \in \Omega^c : |\boldsymbol{x}| < R\}$ for sufficiently large $R > 0$ to ensure that $\overline{\Omega}$ is contained in the disc of radius R.

We begin with the **impedance problem** in the E- or H-mode, see Section 2.11:

Determine $u \in C^2(\Omega^c) \cap C(\overline{\Omega^c})$ such that

$$\Delta u \,+\, k^2 u \;=\; 0 \text{ in } \Omega^c\,, \tag{5.1a}$$

and

$$\frac{\partial u}{\partial n} \,+\, \lambda u \;=\; g \text{ on } \partial\Omega\,, \tag{5.1b}$$

and u satisfies the radiation condition

$$\frac{\partial u}{\partial r} \,-\, iku \;=\; \mathcal{O}\left(\frac{1}{r^{3/2}}\right) \text{ for } r = |\boldsymbol{x}| \to \infty. \tag{5.1c}$$

For a perfectly conducting antenna, the E-mode leads to the **Dirichlet problem** and the H-mode to the **Neumann problem** in the exterior of Ω. Since the Neumann problem is a special case of the impedance problem (set $\lambda = 0$ in (5.1a)–(5.1c)) it remains to formulate the Dirichlet problem:

Determine $u \in C^2(\Omega^c) \cap C(\overline{\Omega^c})$ such that

$$\Delta u \,+\, k^2 u \;=\; 0 \text{ in } \Omega^c\,, \tag{5.2a}$$

and

$$u \;=\; g \text{ on } \partial\Omega\,, \tag{5.2b}$$

and u satisfies the radiation condition (5.1c).

The formulation of the impedance boundary value problem needs a (technical) explanation. Since we did not require that the solution is continuously differentiable up to the boundary, we must explain in which sense the boundary condition has to be understood. The notion of "parallel curves" is very illustrative (see, for example, [99]:

Definition 5.1. *Let $g \in C(\partial\Omega)$. The function $u \in C^1(\Omega^c)$ is said to assume the Neumann boundary value $\partial u/\partial n = g$ on $\partial\Omega$* **in the sense of uniform convergence along the normal** *if*

$$\lim_{t\to 0+} \big|\boldsymbol{n}(\boldsymbol{x}) \cdot \nabla u\big(\boldsymbol{x} + t\boldsymbol{n}(\boldsymbol{x})\big) - g(\boldsymbol{x})\big| \;=\; 0 \quad \textit{uniformly in } \boldsymbol{x} \in \partial\Omega\,. \tag{5.3}$$

Introducing the notation $u_t(\boldsymbol{x}) := u\big(\boldsymbol{x}+t\boldsymbol{n}(\boldsymbol{x})\big)$ and $\nabla u_t(\boldsymbol{x}) := \nabla u\big(\boldsymbol{x}+t\boldsymbol{n}(\boldsymbol{x})\big)$ for sufficiently small $|t|$ we can formulate the boundary condition $\partial u/\partial n = g$ on $\partial\Omega$ as

$$\|\boldsymbol{n} \cdot \nabla u_t - g\|_{C(\partial\Omega)} \;\longrightarrow\; 0 \text{ as } t \to 0+\,, \tag{5.4}$$

where $\|g\|_{C(\partial\Omega)} := \max_{\boldsymbol{x}\in\partial\Omega} |g(\boldsymbol{x})|$ denotes the canonical norm in $C(\partial\Omega)$. The impedance boundary condition (5.1b) may now be understood as

$$\lim_{t\to 0+} \|\boldsymbol{n} \cdot \nabla u_t + \lambda u - g\|_{C(\partial\Omega)} \;=\; 0\,. \tag{5.5}$$

For brevity, we use the notation

$$u(\boldsymbol{x})\big|_{\pm} := \lim_{t\to 0\pm} u_t(\boldsymbol{x})\,, \tag{5.6a}$$

$$\nabla u(\boldsymbol{x})\big|_{\pm} := \lim_{t\to 0\pm} \nabla u_t(\boldsymbol{x})\,, \tag{5.6b}$$

$$\frac{\partial u(\boldsymbol{x})}{\partial n}\bigg|_{\pm} := \lim_{t\to 0\pm} \boldsymbol{n}(\boldsymbol{x}) \cdot \nabla u_t(\boldsymbol{x})\,. \tag{5.6c}$$

Before we discuss existence and uniqueness of these boundary value problems we recall the basic representation theorems for solutions of the Helmholtz equation. For a proof we refer to [29]. The two dimensional fundamental solution of the Helmholtz equation is given by

$$\Phi(\boldsymbol{x}, \boldsymbol{y}) \;:=\; \frac{i}{4}\, H_0^{(1)}(k\,|\boldsymbol{x} - \boldsymbol{y}|) \quad \text{for } \boldsymbol{x} \neq \boldsymbol{y}\,, \tag{5.7}$$

where $H_0^{(1)}$ denotes the Hankel function of the first type and order 0.

Theorem 5.2. *(Representation Theorem)*

Let $\Omega \subset \mathbb{R}^2$ be an open bounded set with exterior $\Omega^c := \mathbb{R}^2 \setminus \overline{\Omega}$, and with the unit normal vector $\boldsymbol{n}(\boldsymbol{x})$ be directed into the exterior Ω^c. Let $k \in \mathbb{C}$ with $\operatorname{Re} k \geq 0$ and $\operatorname{Im} k \geq 0$.

(a) If $u \in C^2(\Omega) \cap C(\overline{\Omega})$ is a solution of the Helmholtz equation $\Delta u + k^2 u = 0$ in Ω which possesses Neumann boundary data in the sense of (5.4), then we have

$$\int_{\partial\Omega} \left\{ \Phi(\boldsymbol{x}, \boldsymbol{y})\, \frac{\partial u(\boldsymbol{y})}{\partial n} - u(\boldsymbol{y})\, \frac{\partial \Phi(\boldsymbol{x}, \boldsymbol{y})}{\partial n(\boldsymbol{y})} \right\} d\ell(\boldsymbol{y}) \;=\; \begin{cases} u(\boldsymbol{x}), & \boldsymbol{x} \in \Omega, \\ 0, & \boldsymbol{x} \in \Omega^c. \end{cases} \tag{5.8a}$$

(b) If $u \in C^2(\Omega^c) \cap C(\overline{\Omega^c})$ be a solution of the Helmholtz equation $\Delta u + k^2 u = 0$ in Ω^c which possesses Neumann boundary data in the sense of (5.4) and satisfies the radiation condition (5.1c), then we have

$$\int\limits_{\partial\Omega} \left\{ \Phi(\boldsymbol{x}, \boldsymbol{y}) \frac{\partial u(\boldsymbol{y})}{\partial n} - u(\boldsymbol{y}) \frac{\partial \Phi(\boldsymbol{x}, \boldsymbol{y})}{\partial n(\boldsymbol{y})} \right\} d\ell(\boldsymbol{y}) = \begin{cases} -u(\boldsymbol{x}), & \boldsymbol{x} \in \Omega^c, \\ 0, & \boldsymbol{x} \in \Omega. \end{cases} \tag{5.8b}$$

Remark: Actually, the proof in [29] is only given for $u \in C^2(\Omega) \cap C^1(\overline{\Omega})$ or $u \in C^2(\Omega^c) \cap C^1(\overline{\Omega^c})$, respectively. But an application of this result in the region bounded by the curves $\{\boldsymbol{x} + t\sigma \boldsymbol{n}(\boldsymbol{x}) : \boldsymbol{x} \in \partial\Omega\}$ where $\sigma = +1$ or -1 respectively leads to the desired result when t tends to zero. For more details on the concept of parallel curves we refer the reader to [74] and [99].

This representation theorem has several important implications. First, we conclude from the analyticity of the kernel in (5.8a) that every C^2- solution of the Helmholtz equation is even analytic. Second, from the asymptotic behaviour of the Hankel function (cf. (2.68a), (2.68b)) we find the following representation of the far field pattern:

Theorem 5.3. *Let $u \in C^2(\Omega^c) \cap C(\overline{\Omega^c})$ be a radiating solution of the Helmholtz equation in Ω^c which assumes Neumann boundary data in sense of (5.4). Then*

$$u(\boldsymbol{x}) = \frac{e^{ikr}}{\sqrt{r}} \left\{ u_\infty(\hat{\boldsymbol{x}}) + \mathcal{O}\left(\frac{1}{r}\right) \right\}, \quad r = |\boldsymbol{x}| \to \infty, \tag{5.9}$$

uniformly in $\hat{\boldsymbol{x}} = \boldsymbol{x}/r \in S^1$. The **far field pattern** *u_∞ has the representation*

$$u_\infty(\hat{\boldsymbol{x}}) = \gamma \int\limits_{\partial\Omega} \left\{ u(\boldsymbol{y}) \frac{\partial}{\partial n(\boldsymbol{y})} e^{-ik\hat{\boldsymbol{x}}\cdot\boldsymbol{y}} - e^{-ik\hat{\boldsymbol{x}}\cdot\boldsymbol{y}} \frac{\partial u(\boldsymbol{y})}{\partial n} \right\} d\ell(\boldsymbol{y}), \quad \hat{\boldsymbol{x}} \in S^1, \tag{5.10}$$

with $\gamma = \exp(i\pi/4)/\sqrt{8\pi k}$. Moreover, the far field pattern u_∞ is an analytic function on the unit circle.

5.2 Rellich's Lemma and Uniqueness

The following lemma describes a fundamental difference between solutions of the Helmholtz equation $\Delta u + k^2 u = 0$ for real $k > 0$ and those of the Laplace equation $\Delta u = 0$. It is well known that there exist solutions of the Laplace equation which decay as $1/r^m$ for $r \to \infty$ for any $m \in \mathbb{N}$ (e.g. the function $u(r, \phi) = r^{-m} \cos(m\phi)$). This situation cannot occur for solutions of the Helmholtz equation for *real and positive k* as was shown by Rellich [113]:

Lemma 5.4. *(Rellich)*
Let $k > 0$ be real and assume that u is a solution of the Helmholtz equation $\Delta u + k^2 u = 0$ in the region $\{\boldsymbol{x} \in \mathbb{R}^2 : |\boldsymbol{x}| > R\}$ for some $R > 0$. Assume that

$$\lim_{r \to \infty} \int\limits_{|\boldsymbol{x}|=r} |u(\boldsymbol{x})|^2 \, d\ell \;=\; 0 \,. \tag{5.11}$$

Then u vanishes for $|\boldsymbol{x}| > R$.

For a proof we refer to [30], Lemma 2.11. This lemma implies that, for real $k > 0$, the mapping $u \mapsto u_\infty$ is one-to-one: $u_\infty = 0$ implies that $u(\boldsymbol{x}) = \mathcal{O}\big(1/r^{3/2}\big)$. Thus condition (5.11) of Rellich's lemma is satisfied, and u must vanish.

Before we turn to the question of uniqueness of solutions for the boundary value problems, we state another implication (cf. [30], Theorem 2.12):

Theorem 5.5. *Let $u \in C^2(\Omega^c) \cap C(\overline{\Omega^c})$ be a radiating solution of the Helmholtz equation in Ω^c which possesses Neumann boundary data in sense of (5.4). Assume that*

$$\operatorname{Im} \int\limits_{\partial\Omega} \overline{u} \frac{\partial u}{\partial n} \, d\ell \;\leq\; 0 \,. \tag{5.12}$$

Then $u = 0$ in Ω^c.

Proof: The binomial theorem yields

$$\int\limits_{|x|=R} \left[\left| \frac{\partial u}{\partial r} \right|^2 + k^2 |u|^2 \right] d\ell \;-\; 2k \operatorname{Im} \int\limits_{|x|=R} \overline{u} \frac{\partial u}{\partial r} \, d\ell \;=\; \int\limits_{|x|=R} \left| \frac{\partial u}{\partial r} - iku \right|^2 d\ell \,, \tag{5.13}$$

and this tends to zero by the radiation condition as $R \to \infty$. By Green's first theorem applied in the region $D_R = \{x \in \Omega^c : |x| < R\}$ we have that

$$\int\limits_{\partial\Omega} \overline{u} \frac{\partial u}{\partial n} \, d\ell \;-\; \int\limits_{|x|=R} \overline{u} \frac{\partial u}{\partial r} \, d\ell \;=\; \iint\limits_{D_R} \left[|\nabla u|^2 - k^2 |u|^2 \right] dx \tag{5.14}$$

and thus

$$\operatorname{Im} \int\limits_{|x|=R} \overline{u} \frac{\partial u}{\partial r} \, d\ell \;=\; \operatorname{Im} \int\limits_{\partial\Omega} \overline{u} \frac{\partial u}{\partial n} \, d\ell \,.$$

Substituting this into (5.13) yields

$$\lim_{R \to \infty} \int\limits_{|x|=R} \left[\left| \frac{\partial u}{\partial r} \right|^2 + k^2 |u|^2 \right] d\ell \;=\; 2k \operatorname{Im} \int\limits_{\partial\Omega} \overline{u} \frac{\partial u}{\partial n} \, d\ell \;\leq\; 0$$

and thus $\lim_{R\to\infty} \int_{|x|=R} |u|^2 d\ell = 0$. Rellich's Lemma yields $u \equiv 0$ outside any ball containing Ω in its interior. Finally, analytic continuation yields that u vanishes everywhere. □

The uniqueness of solutions of the Dirichlet and impedance boundary value problems follows (almost) immediately:

Theorem 5.6. *Let be $k > 0$ and $\lambda \in C(\partial\Omega)$ such that $\operatorname{Im}\lambda \geq 0$ on $\partial\Omega$.*[2] *Then the impedance boundary value problem (5.1a)–(5.1c) has at most one solution.*

Proof: Let u be the difference of two solutions. Then u solves the impedance problem with $g = 0$. Making use of the boundary condition yields

$$\int_{\partial\Omega} \overline{u}\,\frac{\partial u}{\partial n}\,d\ell \;=\; -\int_{\partial\Omega} \lambda\,|u|^2\,d\ell\,.$$

Taking the imaginary part and applying the preceding theorem yields the assertion. □

The Dirichlet problem seems to be even simpler since the integral $\int_{\partial\Omega} \overline{u}\,\frac{\partial u}{\partial n}\,d\ell$ vanishes if, again, u denotes the difference of two solutions. For the Dirichlet problem, however, we have not assumed that the normal derivative exists in the sense of uniform convergence along the normal. Theorem 5.5 is therefore not directly applicable. This problem can be overcome by applying a regularity result as in [29], Theorem 3.27, or proving the identity (5.14) (for the case $u \equiv 0$ on $\partial\Omega$) directly with the following result, originally due to E. Heinz, see [42].

Lemma 5.7. *Let $u \in C^2(\mathbb{R}^2 \setminus \overline{\Omega}) \cap C(\mathbb{R}^2 \setminus \Omega)$ be a solution of the Helmholtz equation in $\mathbb{R}^2 \setminus \overline{\Omega}$ with $u = 0$ on $\partial\Omega$. Define again $D_R := \{x \in \Omega^c : |x| < R\}$ for sufficiently large R. Then $\nabla u \in L^2(\Omega_R)$ and*

$$\iint_{D_R} \left[|\nabla u|^2 - k^2|u|^2\right] dx \;=\; \int_{|x|=R} \frac{\partial u}{\partial r}\,\overline{u}\,d\ell\,.$$

For a proof we refer to [30], Lemma 3.8.

Theorem 5.8. *Let $k > 0$ and $u \in C^2(\Omega^c) \cap C(\overline{\Omega^c})$ be a radiating solution of the Helmholtz equation with $u|_{\partial\Omega} = 0$. Then u has to vanish in Ω^c.*

[2] This condition on $\operatorname{Im}\lambda$ is associated with the physical phenomenon of lossy media.

5.3 Existence by the Boundary Integral Equation Method

In this section we will prove existence of a solution for smooth boundary data λ and g. In contrast to Section 5.4 where we treat L^2–boundary data in this section we restrict ourselves to **Hölder continuous functions**:

Definition 5.9. *Let $\alpha \in (0,1]$ and $G \subset \mathbb{R}^2$ be a set. The space $C^{0,\alpha}(G)$ of bounded, uniformly Hölder continuous functions of order α consists of all bounded functions $f \in C(G)$ for which a constant $c > 0$ exists such that*

$$|f(\boldsymbol{x}) - f(\boldsymbol{y})| \;\leq\; c\,|\boldsymbol{x} - \boldsymbol{y}|^{\alpha} \quad \textit{for all } \boldsymbol{x}, \boldsymbol{y} \in G\,.$$

$C^{0,\alpha}(G)$ is a Banach space with respect to the norm

$$\|f\|_{C^{0,\alpha}(G)} \;:=\; \sup_{\boldsymbol{x}\in G} |f(\boldsymbol{x})| \;+\; \sup_{\boldsymbol{x}\neq\boldsymbol{y}} \frac{|f(\boldsymbol{x}) - f(\boldsymbol{y})|}{|\boldsymbol{x} - \boldsymbol{y}|^{\alpha}}\,.$$

In the integral equation method one searches for the solution in form of a single or double layer potentials. Let again

$$\Phi(\boldsymbol{x}, \boldsymbol{y}) \;:=\; \frac{i}{4}\, H_0^{(1)}(k\,|\boldsymbol{x} - \boldsymbol{y}|) \quad \text{for } \boldsymbol{x} \neq \boldsymbol{y}\,, \tag{5.15}$$

be the fundamental solution of the two dimensional Helmholtz equation. Given an integrable function φ on $\partial\Omega$ the functions

$$u(\boldsymbol{x}) := \int_{\partial\Omega} \varphi(\boldsymbol{y})\, \Phi(\boldsymbol{x}, \boldsymbol{y})\, d\ell(\boldsymbol{y})\,, \quad \boldsymbol{x} \in \mathbb{R}^2 \setminus \partial\Omega\,, \tag{5.16a}$$

$$v(\boldsymbol{x}) := \int_{\partial\Omega} \varphi(\boldsymbol{y})\, \frac{\partial}{\partial n(\boldsymbol{y})} \Phi(\boldsymbol{x}, \boldsymbol{y})\, d\ell(\boldsymbol{y})\,, \quad \boldsymbol{x} \in \mathbb{R}^2 \setminus \partial\Omega\,, \tag{5.16b}$$

are called **single** and **double layer potentials**, respectively, with density φ. They are solutions of the Helmholtz equations and satisfy the radiation condition (5.1c). They are, however, not continuously differentiable when passing through $\partial\Omega$. Their traces are given by the following theorem which we formulate for uniformly Hölder continuous densities φ. These relations are usually referred to as the jump relations for the single and double layer operators.

Theorem 5.10. *Assume that $\varphi \in C^{0,\alpha}(\partial\Omega)$ for some $\alpha \in (0,1]$.*

(a) The single layer potential u can be extended to all of $\mathbb{R}^2$ as a uniformly Hölder continuous function having boundary values given by

$$u(\boldsymbol{x})|_{\pm} \;=\; \int_{\partial\Omega} \varphi(\boldsymbol{y})\, \Phi(\boldsymbol{x}, \boldsymbol{y})\, d\ell(\boldsymbol{y})\,, \quad \boldsymbol{x} \in \partial\Omega\,. \tag{5.17a}$$

Here, $u(\boldsymbol{x})|_{\pm}$ is the limit from the exterior and interior, respectively, as defined in (5.6a)–(5.6c). The integral exists as an improper integral.

(b) *The first derivatives of the single layer potential u can be extended from Ω to $\overline{\Omega}$ and from Ω^c to $\overline{\Omega^c}$ as uniformly Hölder continuous having boundary values*

$$\nabla u(\boldsymbol{x})_\pm \;=\; \mp\frac{1}{2}\varphi(\boldsymbol{x})\,\boldsymbol{n}(\boldsymbol{x}) + \int\limits_{\partial\Omega} \varphi(\boldsymbol{y})\,\nabla_x\Phi(\boldsymbol{x},\boldsymbol{y})\,d\ell(\boldsymbol{y}), \quad \boldsymbol{x}\in\partial\Omega. \tag{5.17b}$$

The integral exists in the sense of the **Cauchy principal value.**[3]

(c) *The double layer potential v can be extended from Ω to $\overline{\Omega}$ and from Ω^c to $\overline{\Omega^c}$ as uniformly continuous functions having boundary values*

$$v(\boldsymbol{x})|_\pm \;=\; \pm\frac{1}{2}\varphi(\boldsymbol{x}) + \int\limits_{\partial\Omega} \varphi(\boldsymbol{y})\,\frac{\partial}{\partial n(\boldsymbol{y})}\Phi(\boldsymbol{x},\boldsymbol{y})\,d\ell(\boldsymbol{y}), \quad \boldsymbol{x}\in\partial\Omega. \tag{5.17c}$$

The integral exists as an improper integral.

(d) *For the normal derivative of the double layer potential v we have*

$$\lim_{t\to 0+} \boldsymbol{n}(\boldsymbol{x})\cdot\big\{\nabla v\big(\boldsymbol{x}+t\,\boldsymbol{n}(\boldsymbol{x})\big) - \nabla v\big(\boldsymbol{x}-t\,\boldsymbol{n}(\boldsymbol{x})\big)\big\} \;=\; 0 \tag{5.17d}$$

uniformly in $\boldsymbol{x}\in\partial\Omega$. Furthermore, there exists $c>0$ with

$$\|u\|_{C^{0,\alpha}(\mathbb{R}^2)} \le c\,\|\varphi\|_{C^{0,\alpha}(\partial\Omega)}, \tag{5.18a}$$

$$\|\nabla u\|_{C^{0,\alpha}(\Omega)} \le c\,\|\varphi\|_{C^{0,\alpha}(\partial\Omega)}, \tag{5.18b}$$

$$\|\nabla u\|_{C^{0,\alpha}(\Omega^c)} \le c\,\|\varphi\|_{C^{0,\alpha}(\partial\Omega)}, \tag{5.18c}$$

$$\|v\|_{C^{0,\alpha}(\Omega)} \le c\,\|\varphi\|_{C^{0,\alpha}(\partial\Omega)}, \tag{5.18d}$$

$$\|v\|_{C^{0,\alpha}(\Omega^c)} \le c\,\|\varphi\|_{C^{0,\alpha}(\partial\Omega)}. \tag{5.18e}$$

For a proof of this theorem we refer to [29], Theorems 2.12, 2.16, 2.17, 2.23.

The integrals which appear in (5.17a)–(5.17c) each define a linear operator in the space of Hölder continuous functions defined on the boundary ∂D. Specifically, we define the boundary operators S, D, and D' by

$$(S\varphi)(\boldsymbol{x}) := \int\limits_{\partial\Omega} \varphi(\boldsymbol{y})\,\Phi(\boldsymbol{x},\boldsymbol{y})\,d\ell(\boldsymbol{y}), \quad \boldsymbol{x}\in\partial\Omega, \tag{5.19a}$$

$$(D\varphi)(\boldsymbol{x}) := \int\limits_{\partial\Omega} \varphi(\boldsymbol{y})\,\frac{\partial}{\partial n(\boldsymbol{y})}\Phi(\boldsymbol{x},\boldsymbol{y})\,d\ell(\boldsymbol{y}), \quad \boldsymbol{x}\in\partial\Omega, \tag{5.19b}$$

$$(D'\varphi)(\boldsymbol{x}) := \int\limits_{\partial\Omega} \varphi(\boldsymbol{y})\,\frac{\partial}{\partial n(\boldsymbol{x})}\Phi(\boldsymbol{x},\boldsymbol{y})\,d\ell(\boldsymbol{y}), \quad \boldsymbol{x}\in\partial\Omega. \tag{5.19c}$$

[3] The notion of the Cauchy principal value extends the ordinary notion of improper integrals to functions on $\mathbb{R}$ or curves which have singularities of order $1/t$. We refer to [137] for an introduction and an application to integral equations.

For C^2–boundaries, these operators are compact:

Theorem 5.11. *The boundary operators S, D and D' are well defined and compact as operators from $C^{0,\alpha}(\partial\Omega)$ into itself.*

Now we turn to the question of existence of solutions to the impedance and Dirichlet boundary problem. We start with the Dirichlet problem and make the assumption that the solution u can be written as combination of a single and double layer potential:

$$u(\boldsymbol{x}) \;=\; \int\limits_{\partial\Omega} \varphi(\boldsymbol{y}) \left[\frac{\partial}{\partial n(\boldsymbol{y})}\Phi(\boldsymbol{x},\boldsymbol{y}) \;-\; ik\Phi(\boldsymbol{x},\boldsymbol{y})\right] d\ell(\boldsymbol{y}), \quad \boldsymbol{x} \in \mathbb{R}^2 \setminus \partial\Omega. \tag{5.20}$$

This form of u satisfies both the Helmholtz equation and the radiation condition. By Theorem 5.10 we see that $u|_+ = g$ on $\partial\Omega$ if and only if the density $\varphi \in C^{0,\alpha}(\partial\Omega)$ solves the boundary integral equation

$$\frac{1}{2}\varphi(\boldsymbol{x}) \;+\; \int\limits_{\partial\Omega} \varphi(\boldsymbol{y}) \left[\frac{\partial}{\partial n(\boldsymbol{y})}\Phi(\boldsymbol{x},\boldsymbol{y}) \;-\; ik\Phi(\boldsymbol{x},\boldsymbol{y})\right] d\ell(\boldsymbol{y}) \;=\; g(\boldsymbol{x}), \quad \boldsymbol{x} \in \partial\Omega, \tag{5.21}$$

or in operator notation

$$\frac{1}{2}\varphi \;+\; D\varphi \;-\; ik\,S\varphi \;=\; g\,. \tag{5.22}$$

This equation is a Fredholm equation of the second kind in $C^{0,\alpha}(\partial\Omega)$ since the operators D and S are compact. By the Theorem of Riesz (Theorem A.40) existence of a solution of this integral equation (5.22) follows once uniqueness is shown. Therefore, let $\varphi \in C^{0,\alpha}(\partial\Omega)$ be the difference of two solutions of (5.22). Then φ solves this integral equation with $g = 0$. Define u by formula (5.20). Then u solves the Helmholtz equation in $\mathbb{R}^2 \setminus \partial\Omega$ and the radiation condition as $r \to \infty$. Furthermore, $u|_+ = \frac{1}{2}\varphi + D\varphi - ik\,S\varphi = 0$ since φ solves (5.22) for $g = 0$. The uniqueness result of Theorem 5.8 then shows that $u = 0$ in Ω^c.

Now we compute the jumps of u and $\partial u/\partial n$ at the boundary $\partial\Omega$ using Theorem 5.10 again:

$$-u|_- \;=\; u|_+ - u|_- \;=\; \varphi \quad\text{and}\quad -\left.\frac{\partial u}{\partial n}\right|_- \;=\; \left.\frac{\partial u}{\partial n}\right|_+ - \left.\frac{\partial u}{\partial n}\right|_- \;=\; ik\varphi\,. \tag{5.23}$$

The trace of the normal derivative has to be understood in the sense of uniform convergence along the normal. Eliminating φ from these equations leads to an impedance boundary condition for u in Ω:

$$\left.\frac{\partial u}{\partial n}\right|_- \;-\; iku|_- \;=\; 0 \quad \text{on } \partial\Omega\,.$$

Now we apply Green's first formula for u and $\overline{u}$ in Ω and make use of the boundary condition. These computations yield

$$\iint\limits_{\Omega} \left[|\nabla u|^2 - k^2 \, |u|^2 \right] dx = \int\limits_{\partial\Omega} \overline{u} \, \frac{\partial u}{\partial n} \, d\ell = ik \int\limits_{\partial\Omega} |u|^2 \, d\ell .$$

Taking the imaginary part gives $u|_- = 0$ and, from (5.23), $\varphi = 0$. Thus we have proven:

Theorem 5.12. *The Dirichlet problem (5.1) has a unique solution for every $k > 0$ and $g \in C^{0,\alpha}(\partial\Omega)$. The solution can be represented in the form (5.20) where $\varphi \in C^{0,\alpha}(\partial\Omega)$ is the unique solution of (5.22).*

Having solved the Dirichlet problem by the integral equation method we can compute the far field pattern u_∞ of the solution by the ansatz and the asymptotic behaviour of the Hankel functions (cf. (2.68a), (2.68b)). This gives

$$u_\infty(\hat{\boldsymbol{x}}) = \gamma \int\limits_{\partial\Omega} \varphi(\boldsymbol{y}) \left[\frac{\partial}{\partial n(\boldsymbol{y})} \mathrm{e}^{-ik\hat{\boldsymbol{x}}\cdot\boldsymbol{y}} - ik \, \mathrm{e}^{-ik\hat{\boldsymbol{x}}\cdot\boldsymbol{y}} \right] d\ell(\boldsymbol{y}) , \quad \hat{\boldsymbol{x}} \in S^1 , \quad (5.24)$$

with $\gamma = \sqrt{\frac{2}{\pi k}} \, \mathrm{e}^{-i\pi/4}$.

The situation is more complicated for the impedance problem. This is due to the fact that the normal derivative of the double layer potential does not exist for Hölder continuous densities, not even in the sense of uniform convergence. It does exist, however, for Hölder continuously differentiable functions. There are different possibilities to introduce the Hölder space $C^{1,\alpha}(\partial\Omega)$ of Hölder continuously differentiable functions. Instead of using local coordinates as, e.g. in [30] (Section 6.3 for surfaces), we prefer the following approach which we will extend to surfaces in $\mathbb{R}^3$ in the next chapter. We begin with the spaces $C_T(\partial\Omega)$ and $C_T^{0,\alpha}(\partial\Omega)$ of **tangential vector fields** defined as

$$C_T(\partial\Omega) := \left\{ \boldsymbol{a} : \partial\Omega \to \mathbb{C}^2 : a_j \in C(\partial\Omega), \ j = 1,2, \ \boldsymbol{a}(\boldsymbol{x}) \cdot \boldsymbol{n}(\boldsymbol{x}) = 0 \text{ on } \partial\Omega \right\},$$
$$C_T^{0,\alpha}(\partial\Omega) := \left\{ \boldsymbol{a} \in C_T(\partial\Omega) : a_j \in C^{0,\alpha}(\partial\Omega), \ j = 1,2 \right\}.$$

We recall that a function $f : \partial\Omega \longrightarrow \mathbb{C}$ is continuously differentiable if there exists a tangential vector field $\operatorname{Grad} f \in C_T(\partial\Omega)$ such that

$$\lim_{t\to 0} \frac{1}{t} \left| f\big(\boldsymbol{x} + t\boldsymbol{a}(\boldsymbol{x})\big) - f(\boldsymbol{x}) - t \operatorname{Grad} f(\boldsymbol{x}) \cdot \boldsymbol{a}(\boldsymbol{x}) \right| = 0 ,$$

for all tangential vector fields $\boldsymbol{a} \in C_T(\partial\Omega)$. The vector field $\operatorname{Grad} f$ is called the **tangential gradient** of f.

Definition 5.13. *Let $\alpha \in (0,1]$. The space $C^{1,\alpha}(\partial\Omega)$ consists of all functions $f \in C(\partial\Omega)$ which are continuously differentiable on $\partial\Omega$ and whose tangential gradients* Grad f, *are Hölder continuous of order α, i.e.* $\operatorname{Grad} f \in C_T^{0,\alpha}(\partial\Omega)$.

The space $C^{1,\alpha}(\partial\Omega)$ is also a Banach space with respect to the canonical norm

$$\|f\|_{C^{1,\alpha}(\partial\Omega)} \;:=\; \|f\|_{C^{0,\alpha}(\partial\Omega)} \;+\; \|\operatorname{Grad} f\|_{C^{0,\alpha}_T(\partial\Omega)} \,.$$

Also, the imbeddings $j : C^{0,\alpha}(\partial\Omega) \longrightarrow C(\partial\Omega)$ and $C^{1,\alpha}(\partial\Omega) \to C^{0,\alpha}(\partial\Omega)$ are compact. For the proofs of these facts we refer to [29, p.38].

For densities $\varphi \in C^{1,\alpha}(\partial\Omega)$ the following holds:

Theorem 5.14. *Let $\varphi \in C^{1,\alpha}(\partial\Omega)$ for some $\alpha \in (0,1]$. Then the double layer potential v and its first derivatives can be continuously extended from Ω to $\overline{\Omega}$ and from Ω^c to $\overline{\Omega^c}$ with boundary values*

$$\partial v/\partial n|_+ \;=\; \partial v/\partial n|_- \,. \tag{5.25}$$

The operator T, defined by $T\varphi := \partial v/\partial n$, i.e.

$$T\varphi(\boldsymbol{x}) \;:=\; \frac{\partial}{\partial n}\int\limits_{\partial\Omega} \varphi(\boldsymbol{y})\,\frac{\partial}{\partial n(\boldsymbol{y})}\Phi(\boldsymbol{x},\boldsymbol{y})\,d\ell(\boldsymbol{y})\,, \quad \boldsymbol{x} \in \partial\Omega, \tag{5.26}$$

is well defined and bounded as an operator from $C^{1,\alpha}(\partial\Omega)$ into $C^{0,\alpha}(\partial\Omega)$. The operator D from (5.19a) is bounded from $C^{0,\alpha}(\partial\Omega)$ into $C^{1,\alpha}(\partial\Omega)$.

For a proof we refer to [29], Theorems 2.23 and 2.30. For the following we observe that we can easily compute the L^2-adjoints of the boundary operators S, D, D' and T. By $\langle\cdot,\cdot\rangle_{L^2(\partial\Omega)}$ we denote the L^2-inner product.

Theorem 5.15. *For an operator K in some function space we denote by $\overline{K}$ the complex conjugate of K, i.e. $\overline{K}\varphi := \overline{K\overline{\varphi}}$. Then we have*

(a) $\langle\overline{D}\varphi,\psi\rangle_{L^2(\partial\Omega)} \;=\; \langle\varphi, D'\psi\rangle_{L^2(\partial\Omega)}$ for all $\psi,\varphi \in C^{0,\alpha}(\partial\Omega)$, i.e. $\overline{D}$ and D' are L^2-adjoint of each other.
(b) $\langle\overline{S}\varphi,\psi\rangle_{L^2(\partial\Omega)} \;=\; \langle\varphi, S\psi\rangle_{L^2(\partial\Omega)}$ for all $\psi,\varphi \in C^{0,\alpha}(\partial\Omega)$, i.e. $\overline{S}$ and S are L^2-adjoint of each other.
(c) $\langle\overline{T}\varphi,\psi\rangle_{L^2(\partial\Omega)} \;=\; \langle\varphi, T\psi\rangle_{L^2(\partial\Omega)}$ for all $\psi,\varphi \in C^{1,\alpha}(\partial\Omega)$, i.e. $\overline{T}$ and T are L^2-adjoint of each other.

Parts (a) and (b) are easily seen by changing the order of integration. For this and for part (c) we refer to [30].

Having established the definitions of these operators we now continue with the impedance boundary value problem (5.2), (5.1c) where we now assume that $\lambda \in C^{0,\alpha}(\partial\Omega)$. We start by looking for solutions in the form of a combination of a single and double layer potential, but one slightly different than for the Dirichlet problem:

$$u(\boldsymbol{x}) \;=\; \int\limits_{\partial\Omega}\left[(\overline{D}D'\varphi)(\boldsymbol{y})\,\frac{\partial}{\partial n(\boldsymbol{y})}\Phi(\boldsymbol{x},\boldsymbol{y}) \;-ik\,\varphi(\boldsymbol{y})\,\Phi(\boldsymbol{x},\boldsymbol{y})\right]d\ell(\boldsymbol{y})\,, \tag{5.27}$$

for $\boldsymbol{x} \in \mathbb{R}^2 \backslash \partial\Omega$ with some density $\varphi \in C^{0,\alpha}(\partial\Omega)$. This form of u again satisfies the Helmholtz equation and the radiation condition. The density of the single layer part is $-ik\varphi$, for the double layer part it is $\overline{D}D'\varphi$ and thus in $C^{1,\alpha}(\partial\Omega)$ by Theorem 5.14. By Theorems 5.10 and 5.14 we see that $\partial u/\partial n|_+ + \lambda\, u|_+ = g$ on $\partial\Omega$ if and only if the density $\varphi \in C^{0,\alpha}(\partial\Omega)$ solves the boundary integral equation

$$T\overline{D}D'\varphi + \frac{ik}{2}\varphi - ik\,D'\varphi + \frac{\lambda}{2}\overline{D}D'\varphi + \lambda\, D\overline{D}D'\varphi - ik\,\lambda\, S\varphi \;=\; g\,. \quad (5.28)$$

This equation looks complicated. But we see that it is also a Fredholm equation of the second kind since the operators $T\overline{D}D'$, D', $\overline{D}D'$, $D\overline{D}D'$, and S are all compact in $C^{0,\alpha}(\partial\Omega)$. Again, we have to prove uniqueness. Let $\varphi \in C^{0,\alpha}(\partial\Omega)$ be a solution of (5.28) for $g = 0$. Define u by (5.27). Then u solves the exterior impedance problem (5.2), (5.1c) with homogeneous boundary data $g = 0$. If uniqueness holds for the boundary value problem itself (e.g. if $\operatorname{Im}\lambda \geq 0$ on $\partial\Omega$ by Theorem 5.6), we conclude that u vanishes in Ω^c. Again we calculate the jumps with Theorems 5.10 and 5.14

$$-u|_- \;=\; u|_+ - u|_- \;=\; \overline{D}D'\varphi \quad \text{and} \quad -\left.\frac{\partial u}{\partial n}\right|_- \;=\; \left.\frac{\partial u}{\partial n}\right|_+ - \left.\frac{\partial u}{\partial n}\right|_- \;=\; ik\,\varphi\,. \quad (5.29)$$

Applying the operator $\overline{D}D'$ to the second equation we arrive at the *nonlocal* boundary condition

$$\overline{D}D'\left(\left.\frac{\partial u}{\partial n}\right|_-\right) - ik\,u|_- \;=\; 0 \quad \text{on}\; \partial\Omega\,.$$

Multiplication with $\partial\overline{u}/\partial n\big|_-$, integration over $\partial\Omega$ and using of the fact that $\overline{D}$ and D' are L^2–adjoint, yields (the traces are taken from the interior)

$$\left\| D'\frac{\partial u}{\partial n}\right\|^2_{L^2(\partial\Omega)} \;=\; ik\int\limits_{\partial\Omega} u\,\frac{\partial\overline{u}}{\partial n}\,d\ell \;=\; ik\iint\limits_{\Omega}\left[|\nabla u|^2 - k^2\,|u|^2\right]dx\,.$$

Taking the real part yields $0 = D'\frac{\partial u}{\partial n} = -ikD'\varphi$ on $\partial\Omega$ and thus also $u|_- = 0$ by (5.29). The ansatz reduces to $u(\boldsymbol{x}) = -ik \int\limits_{\partial\Omega} \varphi(\boldsymbol{y})\,\Phi(\boldsymbol{x},\boldsymbol{y})\,d\ell(\boldsymbol{y})$ and thus $0 = u|_- = -ik\,S\varphi$. The integral equation (5.28) finally shows that $\varphi = 0$. Thus we have proven:

Theorem 5.16. *Suppose that there exists at most one solution for the impedance problem (5.2), (5.1c) for $k > 0$ and impedance $\lambda \in C^{0,\alpha}(\partial\Omega)$ (e.g. if $\operatorname{Im}\lambda \geq 0$ on $\partial\Omega$). Then the integral equation (5.28) has a unique solution for every $g \in C^{0,\alpha}(\partial\Omega)$. Furthermore the solution of the original boundary value problem can be represented in the form (5.27) where $\varphi \in C^{0,\alpha}(\partial\Omega)$ is the unique solution of (5.28).*

Again, having solved the impedance problem by this integral equation method we can compute the far field pattern u_∞ of the solution by the ansatz (5.27) as

$$u_\infty(\hat{\boldsymbol{x}}) \;=\; \gamma \int_{\partial\Omega} \left[(\overline{D}\, D'\varphi)(\boldsymbol{y}) \frac{\partial}{\partial n(\boldsymbol{y})} \mathrm{e}^{-ik\hat{\boldsymbol{x}}\cdot\boldsymbol{y}} \;-\; ik\,\varphi(\boldsymbol{y})\,\mathrm{e}^{-ik\hat{\boldsymbol{x}}\cdot\boldsymbol{y}} \right] d\ell(\boldsymbol{y})\,, \tag{5.30}$$

$\hat{\boldsymbol{x}} \in S^1$, with $\gamma = \sqrt{\frac{2}{\pi k}}\,\mathrm{e}^{-i\pi/4}$.

Remark: For numerical algorithms, one usually works with the "plain" double layer approach for the Dirichlet problem and the single layer ansatz for the impedance problem. Then it is well known that the resulting integral equations are not uniquely solvable if k^2 is an eigenvalue of $-\Delta$ in Ω for Neumann- or Dirichlet boundary conditions, respectively. But these eigenvalues are discrete and numerical experiments showed that no numerical problems occur as long as one does not approach these eigenvalues too closely.

We finish this section by introducing the **solution operator** and the **far field operator** of the Dirichlet and impedance boundary problems.

Theorem 5.17. *(a) Let $A \subset \Omega^c$ be any compact subset of Ω^c and $\lambda \in C^{0,\alpha}(\partial\Omega)$. Then the solution operators $\mathcal{L}_D : C^{0,\alpha}(\partial\Omega) \longrightarrow C(A)$ and $\mathcal{L}_I : C^{0,\alpha}(\partial\Omega) \longrightarrow C(A)$ for the Dirichlet and impedance boundary value problems, respectively, are defined by $g \mapsto u|_A$, where u satisfies the Dirichlet or impedance boundary value problem (5.1) or (5.2), (5.1c), respectively. They are well-defined and bounded.*

(b) The far field operators $\mathcal{K}_D : C^{0,\alpha}(\partial\Omega) \longrightarrow C(S^1)$ and $\mathcal{K}_I : C^{0,\alpha}(\partial\Omega) \longrightarrow C(S^1)$ for the Dirichlet and impedance boundary value problems, respectively, are defined by $g \mapsto u_\infty$, where u satisfies the Dirichlet or impedance boundary value problem (5.1) or (5.2), (5.1c), respectively. They are well defined and bounded.

Proof: The representations (5.20) and (5.27) define operators from $C^{0,\alpha}(\partial\Omega)$ into $C(A)$ which are bounded. Since the equations (5.22) and (5.28) define isomorphisms in $C^{0,\alpha}(\partial\Omega)$ part (a) follows. The far field patterns are given by (5.24) and (5.30), respectively, which also define bounded operators from $C^{0,\alpha}(\partial\Omega)$ into $C(S^1)$. This proves the theorem. □

5.4 L^2–Boundary Data

In many applications it is important to allow the boundary data g or the impedance λ to have "jumps" or even more complicated discontinuities. For the important question of existence of solutions of antenna problems (see Chapter 3) it is useful to take, e.g., $g \in L^2(\partial\Omega)$. Then the meaning of the

Dirichlet and Neumann boundary condition has to be clarified. There are a number of equivalent definitions (variational formulations, parallel curves). We refer to [66], [9] for a comprehensive study of these notions. In this chapter we follow an approach which is based on the notion of parallel curves which we have introduced already to overcome the difficulty with the differentiability of u up to the boundary. It is very natural to replace the norm $\|\cdot\|_{C(\partial\Omega)}$ in (5.5) by the L^2−norm if the boundary data is to be taken in $L^2(\partial\Omega)$. Then it can be shown that all of the uniqueness- and existence results of Sections 5.2 and 5.3 carry over. We will only sketch the approach.

Our final goal is to show that the solution operators $\mathcal{L}_D$, $\mathcal{L}_I$ and the far field operators $\mathcal{K}_D$, $\mathcal{K}_I$ of Theorem 5.17 possess extensions to bounded operators from $L^2(\partial\Omega)$ into $C(A)$ and $C(S^1)$, respectively.

First, we formulate the Dirichlet and impedance problems for $g \in L^2(\partial\Omega)$ and $\lambda \in L^\infty(\partial\Omega)$. We recall the definitions of $u_t(\boldsymbol{x}) := u\big(\boldsymbol{x} + t\,\boldsymbol{n}(\boldsymbol{x})\big)$ and $\nabla u_t(\boldsymbol{x}) := \nabla u\big(\boldsymbol{x}+t\,\boldsymbol{n}(\boldsymbol{x})\big)$ for sufficiently small $|t|$. We start with the *Dirichlet problem*:

Determine $u \in C^2(\Omega^c)$ such that

$$\Delta u + k^2 u = 0 \quad \text{in } \Omega^c, \tag{5.31a}$$

$$\|u_t - g\|_{L^2(\partial\Omega)} \longrightarrow 0 \quad \text{as } t \to 0+, \tag{5.31b}$$

and u satisfies the radiation condition

$$\frac{\partial u}{\partial r} - iku = \mathcal{O}\left(\frac{1}{r^{3/2}}\right) \quad \text{as } r = |\boldsymbol{x}| \to \infty \tag{5.31c}$$

uniformly in $\hat{\boldsymbol{x}} := \boldsymbol{x}/r$.

The *impedance problem* is to determine $u \in C^2(\Omega^c)$ such that

$$\Delta u + k^2 u = 0 \quad \text{in } \Omega^c, \tag{5.32a}$$

and

$$\|\boldsymbol{n}\cdot\nabla u_t + \lambda u_t - g\|_{L^2(\partial\Omega)} \longrightarrow 0 \quad \text{as } t \to 0+, \tag{5.32b}$$

and u satisfies the radiation condition (5.31c).

In the following, once again we write $u = g$ on $\partial\Omega$ and $\partial u/\partial n + \lambda u = g$ on $\partial\Omega$ for (5.31b) and (5.32b), respectively, and say that the boundary data exist in the L^2-sense.

Then it can be shown that Theorem 5.5 holds if both, the Dirichlet boundary data u and the Neumann boundary data $\partial u/\partial n$ exist in the L^2−sense. Also, the uniqueness Theorems 5.6 and 5.8 can be carried over to the L^2−case.

For the question of existence we first recall that the single and double layer potentials u and v from (5.16a), (5.16b) exist also for $\varphi \in L^2(\partial\Omega)$. The operators S, D, D' and its L^2−adjoints are compact operators in $C^{0,\alpha}(\partial\Omega)$ as we have seen in the previous section. By the following theorem of Lax (see [57, 30]) we see that they are also bounded in $L^2(\partial\Omega)$:

Theorem 5.18. *Assume that $T : X_1 \longrightarrow X_2$ is a bounded operator between normed spaces X_1 and X_2. Let $\langle \cdot, \cdot \rangle$ be inner products in X_1 and X_2. We denote the corresponding norms by $\|x\|_{H_j} := \sqrt{\langle x, x \rangle}$.*[4] *Assume, furthermore, that the adjoint $T^* : X_2 \longrightarrow X_1$ with respect to $\langle \cdot, \cdot \rangle$ is also bounded in the norms of X_1 and X_2. Then T and T^* can be uniquely extended to bounded operators $\tilde{T} : H_1 \longrightarrow H_2$ and $\tilde{T}^* : H_2 \longrightarrow H_1$, respectively, where H_j, $j = 1, 2$, is the completion of X_j with respect to $\|\cdot\|_{H_j}$.*

If T and T^* are isomorphisms from X_1 onto X_2 and X_2 onto X_1, respectively, then also $\tilde{T}$ and $\tilde{T}^*$ are isomorphisms between H_1 and H_2.

Proof: The boundedness of T and T^* with respect to $\|\cdot\|_{H_j}$ has been shown in, e.g., [30], Theorem 3.5. Therefore, the existence of the bounded extensions $\tilde{T} : H_1 \longleftrightarrow H_2 : \tilde{T}^*$ are guaranteed.

Assume that T and T^* are isomorphisms between X_1 and X_2. Then T^{-1} and $\left(T^*\right)^{-1}$ satisfy the assumptions of the theorem as well and, therefore, can be extended to bounded operators between H_1 and H_2. It is easy to see that these extensions are the inverses of $\tilde{T}$ and $\tilde{T}^*$, respectively. □

Kersten [62] used this theorem to extend the jump conditions (5.17) for single- and double layers with L^2- densities. They hold in the following sense:

$$\|u_{\pm t} - S\varphi\|_{L^2(\partial\Omega)} \longrightarrow 0, \quad t \to 0+ \tag{5.33a}$$

$$\|\boldsymbol{n} \cdot \nabla u_{\pm t} \pm \varphi - D'\varphi\|_{L^2(\partial\Omega)} \longrightarrow 0, \quad t \to 0+ \tag{5.33b}$$

$$\|v_{\pm t} \mp \varphi - D\varphi\|_{L^2(\partial\Omega)} \longrightarrow 0, \quad t \to 0+, \tag{5.33c}$$

$$\|\boldsymbol{n} \cdot \nabla v_{+t} - \boldsymbol{n} \cdot \nabla v_{-t}\|_{L^2(\partial\Omega)} \longrightarrow 0, \quad t \to 0+\,. \tag{5.33d}$$

With these tools it is possible to carry over the integral equation method of Section 5.3 to the case where $g \in L^2(\partial\Omega)$ and $\lambda \in L^\infty(\partial\Omega)$. In particular, the integral equations (5.22) and (5.28) are uniquely solvable (the latter provided $\operatorname{Im}\lambda(\boldsymbol{x}) \geq 0$ almost everywhere on $\partial\Omega$) for every $g \in L^2(\partial\Omega)$. We summarize the results in the following theorem.

Theorem 5.19. *Let be $\lambda \in L^\infty(\partial\Omega)$ with $\operatorname{Im}\lambda(\boldsymbol{x}) \geq 0$ almost everywhere on $\partial\Omega$. Then the Dirichlet and the impedance boundary value problems are uniquely solvable for every $g \in L^2(\partial\Omega)$. The solution operators $\mathcal{L}_D$ and $\mathcal{L}_I$ and the far field operators $\mathcal{K}_D$ and $\mathcal{K}_I$ from Theorem 5.17 possess extensions to linear bounded operators from $L^2(\partial\Omega)$ into $C(A)$ and from $L^2(\partial\Omega)$ into $C(S^1)$, respectively.*

Moreover, the operators $\mathcal{K} = \mathcal{K}_D$ and $\mathcal{K} = \mathcal{K}_I$ are one-to-one with dense ranges in $C(S^1)$, and the images $\mathcal{K}g$, $g \in L^2(\partial\Omega)$, are analytic functions on the unit circle.

[4] Our notation does not distinguish between the inner product or norms in X_1 and X_2.

Proof: We sketch only the proofs of the last assertions. Injectivity is again an immediate consequence of Rellich's Lemma 5.4 and analytic continuation. The fact that far field patterns $\mathcal{K}h$ are analytic functions has been noticed before in Theorem 5.9. For the denseness of the ranges of $\mathcal{K}_D$ and $\mathcal{K}_I$ we recall the **approximation theorem of Weierstrass**. It states that for every function $f \in C(S^1)$ and every $\epsilon > 0$ there exists $N \in \mathbb{N}$ and $a_{-N}, \ldots, a_N \in \mathbb{C}$ such that

$$\left\| \sum_{n=-N}^{N} a_n \psi_n - f \right\|_{C(S^1)} \leq \epsilon ,$$

where $\psi_n(\theta) := \exp(in\theta)$, $n \in \mathbb{Z}$. Here we have identified $C(S^1)$ with the space of continuous periodic functions on $[0, 2\pi]$. Now we observe that $f_N := \sum_{n=-N}^{N} a_n \psi_n$ is the far field pattern of

$$u_N(r, \theta) := \sqrt{\frac{\pi k}{2}} \sum_{n=-N}^{N} a_n H_n^{(1)}(kr)\, \mathrm{e}^{in(\theta+\pi/2)+i\pi/4} ,$$

where $H_n^{(1)}$ denotes the **Hankel function of first kind** and order n (see Section 2.13). Defining g by $g := u_N|_{\partial\Omega}$ for the Dirichlet problem and $g := \left(\partial u_N/\partial n + \lambda u_N\right)\big|_{\partial\Omega}$ for the impedance problem yields $\mathcal{K}g = f_N$ and thus $\|\mathcal{K}g - f\|_{C(S^1)} \leq \epsilon$. □

Example 5.20. In this example we want to demonstrate the use of parallel curves for the simple, but important, example of a boundary value problem where Ω is a disk of radius R with boundary $\Gamma_R = \{\boldsymbol{x} \in \mathbb{R}^2 : |\boldsymbol{x}| = R\}$. First, expand the boundary function $g \in L^2(\Gamma_R)$ in a Fourier series

$$g(\theta) = \sum_{n\in\mathbb{Z}} a_n \,\mathrm{e}^{in\theta} , \quad 0 \leq \theta \leq 2\pi , \tag{5.34}$$

where the convergence is understood in the L^2–sense i.e.,

$$\int_0^{2\pi} \left| g(\theta) - \sum_{|n|\leq N} a_n \,\mathrm{e}^{in\theta} \right|^2 d\theta \longrightarrow 0 \quad \text{as } N \to \infty .$$

Formally, the solution of the exterior Dirichlet problem (5.2) is given by

$$u(r, \theta) = \sum_{n\in\mathbb{Z}} \frac{a_n}{H_n^{(1)}(kR)} H_n^{(1)}(kr)\, \mathrm{e}^{in\theta} , \quad r > R ,\ 0 \leq \theta \leq 2\pi . \tag{5.35}$$

It is the aim of this example to show that this form of u is indeed the weak solution of the exterior Dirichlet problem in the sense introduced above.

First, we note that

$$u_N(r, \theta) = \sum_{|n|\leq N} \frac{a_n}{H_n^{(1)}(kR)} H_n^{(1)}(kr)\, \mathrm{e}^{in\theta} , \quad r > R ,\ 0 \leq \theta \leq 2\pi ,$$

is a classical solution of the exterior Dirichlet boundary value problem with boundary data

$$g_N(\theta) \;=\; \sum_{|n|\le N} a_n\, e^{in\theta}\,, \quad 0 \le \theta \le 2\pi\,.$$

Furthermore, using the asymptotic behaviour of the Hankel functions for large orders (see [1]),

$$H_n^{(1)}(t) \;=\; \frac{(n-1)!}{\pi i}\left(\frac{2}{t}\right)^n \big[1 \,+\, \mathcal{O}(1/n)\big] \quad \text{as } n \to \infty\,, \tag{5.36}$$

uniformly on compact subset of $(0,\infty)$, we see that for $R < a \le r \le b$,

$$|u_N(r,\theta)| \le \sum_{|n|\le N} \frac{|a_n|\,|H_n^{(1)}(kr)|}{|H_n^{(1)}(kR)|} \le c \sum_{|n|\le N} |a_n| \left(\frac{R}{r}\right)^{|n|}$$
$$\le c \sum_{|n|\le N} |a_n| \left(\frac{R}{a}\right)^{|n|} \le c \left(\sum_{|n|\le N} |a_n|^2\right)^{1/2} \left(\sum_{|n|\le N} \left(\frac{R}{a}\right)^{2|n|}\right)^{1/2}$$

for all $\theta \in [0,2\pi]$ and $R < a \le r \le b$. Therefore, the series for u converges uniformly on compact subsets of $\{\boldsymbol{x} \in \mathbb{R}^2 : |\boldsymbol{x}| > R\}$. The same arguments hold also for any derivative of u. From this and the continuous dependence result for classical solutions (see Theorem 5.17) we conclude that u is a classical solution of the exterior Dirichlet problem in the region $\{\boldsymbol{x} \in \mathbb{R}^2 : |\boldsymbol{x}| > R+\epsilon\}$ with boundary data $u|_{\Gamma_{R+\epsilon}}$ for every $\epsilon > 0$. In particular, this implies that $u \in C^2(\Omega^c)$, that it satisfies the Helmholtz equation (5.31a) and the radiation condition (5.31c). It remains to show that it satisfies the boundary condition (5.31b) as well. But

$$u_t(\theta) \;=\; u(R+t,\theta) \;=\; \sum_{n\in\mathbb{Z}} a_n \frac{H_n^{(1)}\big(k(R+t)\big)}{H_n^{(1)}(kR)}\, e^{in\theta}\,,$$

so that

$$\|u_t - g\|^2_{L^2(\Gamma_R)} \;=\; 2\pi \sum_{n\in\mathbb{Z}} |a_n|^2 \left|\frac{H_n^{(1)}\big(k(R+t)\big)}{H_n^{(1)}(kR)} \;-\; 1\right|^2 .$$

We note that, for every fixed $n \in \mathbb{Z}$,

$$\left|\frac{H_n^{(1)}\big(k(R+t)\big)}{H_n^{(1)}(kR)} \;-\; 1\right|^2 \longrightarrow\; 0\,, \quad \text{as } t \to 0\,, \tag{5.37a}$$

and there exists $c > 0$ with

$$\left|\frac{H_n^{(1)}(k(R+t))}{H_n^{(1)}(kR)}\right| \leq c\left|\frac{R}{R+t}\right|^{|n|} \leq c$$

i.e.

$$\left|\frac{H_n^{(1)}(k(R+t))}{H_n^{(1)}(kR)} - 1\right| \leq c+1 \tag{5.37b}$$

for all $t \in [0,1]$ and $n \in \mathbb{Z}$.

These two properties yield convergence $\|u_t - g\|_{L^2(\Gamma_R)} \to 0$ as $t \to 0$ by standard arguments. Indeed, let $\epsilon > 0$ and choose N so large such that

$$\sum_{|n|\geq N+1} |a_n|^2 \leq \frac{\epsilon}{4\pi\,(c+1)^2}\,.$$

Then we split

$$\begin{aligned}\|u_t - g\|^2_{L^2(\Gamma_R)} \leq 2\pi \sum_{|n|\leq N} |a_n|^2 \left|\frac{H_n^{(1)}(k(R+t))}{H_n^{(1)}(kR)} - 1\right|^2 \\ + 2\pi\,(c+1)^2 \sum_{|n|\geq N+1} |a_n|^2\end{aligned}$$

for all $t > 0$. Now we choose $\delta > 0$ such that

$$\sum_{|n|\leq N} |a_n|^2 \left|\frac{H_n^{(1)}(k(R+t))}{H_n^{(1)}(kR)} - 1\right|^2 \leq \frac{\epsilon}{4\pi} \quad \text{for all } 0 < t \leq \delta\,.$$

Then $\|u_t - g\|^2_{L^2(\Gamma_R)} \leq \epsilon$ for all $0 < t \leq \delta$ which proves that the boundary condition (5.31b) is satisfied in the L^2–sense.

We finish this example with the remarks that the solution of the exterior impedance problem for the circle with constant impedance $\lambda \in \mathbb{C}$ with $\operatorname{Im}\lambda \geq 0$ is given by

$$u(r,\theta) = \sum_{n\in\mathbb{Z}} \frac{a_n}{k\,(H_n^{(1)})'(kR) + \lambda\, H_n^{(1)}(kR)}\, H_n^{(1)}(kr)\, e^{in\theta}\,, \tag{5.38}$$

for $r > R$, $0 \leq \theta \leq 2\pi$, and that the far field patterns u_∞ for the Dirichlet and impedance problem are given by

$$u_\infty(\theta) = \gamma \sum_{n\in\mathbb{Z}} \frac{a_n}{H_n^{(1)}(kR)}\, e^{in(\theta-\pi/2)}\,, \quad 0 \leq \theta \leq 2\pi\,, \quad \text{and} \tag{5.39a}$$

$$u_\infty(\theta) = \gamma \sum_{n\in\mathbb{Z}} \frac{a_n}{k\,(H_n^{(1)})'(kR) + \lambda\, H_n^{(1)}(kR)}\, e^{in(\theta-\pi/2)}\,, \tag{5.39b}$$

respectively, where $\gamma = \sqrt{2/(\pi k)}\exp(-i\pi/4)$.

As we see immediately from the definition, the impedance boundary value problem reduces to the Neumann problem for $\lambda = 0$. Moreover, from the integral equation (5.28) we observe that the solution $\varphi = \varphi_\lambda$ of (5.28) converges to the solution φ_0 corresponding to the Neumann problem as λ tends to zero. Indeed, we may rewrite (5.28) in the form

$$A\varphi_\lambda + \lambda B\varphi_\lambda = g, \quad \text{i.e.} \quad \left[I + A^{-1}\lambda B\right]\varphi_\lambda = A^{-1}g,$$

with

$$A = \frac{ik}{2}I - ikD' + T\overline{D}D' \quad \text{and} \quad B = \frac{1}{2}\overline{D}D' + D\overline{D}D' - ikS.$$

It is easy to see that the operator $A^{-1}\lambda B$ tends to zero in the operator norm (either in the Hölder space $C^{0,\alpha}(\partial\Omega)$ or in $L^2(\partial\Omega)$) as λ tends to zero. An application of the perturbation lemma (Theorem A.39 of the Appendix) yields an estimate of the form

$$\|\varphi_\lambda - \varphi_0\| \leq c\,\|\lambda\| \,.$$

On the other hand, writing the impedance boundary condition in the form

$$-\epsilon\frac{\partial u_\epsilon}{\partial n} + u_\epsilon = g \quad \text{on } \partial\Omega,$$

one can ask the same question of convergence as ϵ tends to zero. Note that, in contrast to the case $\lambda \to 0$ in (5.1b), the case $\epsilon \to 0$ leads to a **singular perturbation problem** since the solution u_ϵ corresponding to $\epsilon > 0$ is more regular than the solution corresponding to $\epsilon = 0$. Nevertheless, it has been shown in [65] that u_ϵ converges in $L^2(\partial\Omega)$ to u_0 for every $g \in L^2(\partial\Omega)$. Furthermore, rates of convergence can be given which depend on the smoothness of the boundary data g.

5.5 Numerical Methods

We now turn to a description of some numerical methods for computing the far field patterns of antenna problems for given boundary data. Numerical methods for solving boundary value problems in exterior domains fall into (at least) four groups.

The integral equations derived in the last section could be solved either by **Nyström methods** or by **boundary element methods**. In Subsection 5.5.1 we will recall the Nyström method (see the earlier Subsection 4.5.2), but now applied to integrals with *weakly singular kernels*. This provides a very accurate and fast numerical algorithm provided the data are smooth. In particular, if the boundary $\partial\Omega$, the impedance, and the feeding g are analytic the method is known to converge with exponential order of convergence (see Subsection 5.5.1.

In Subsection 5.5.2 we present a rather simple method based on the observation that the solution of the antenna problem for circular cross sections can be expressed in a series of circular wave functions. Although this is not true for the non-circular case we still know that every solution can approximated (in a sense to be made clear later) by sums of circular wave functions. The coefficients of the sum are determined by the boundary condition and leads to least squares problems.

A variety of **finite element** and **finite difference methods** have been employed also for exterior boundary value problems. One class of methods uses expanding grids (cf. [77, 43]). In these methods R is taken sufficiently large that a simple boundary condition on the circle Γ_R of radius R (e.g. $\partial u/\partial r - iku = 0$ on Γ_R) generates an accurate approximation to u. Methods of this type do not provide direct approximations to the far-field pattern of the solution, although far-field patterns can be deduced using the near-field. At the end of this section we will describe one such method which has received much attention.

A forth class of methods are sometimes called **hybrid methods**. In order to avoid expanding grids in finite element or finite difference methods, Masmoudi [95] suggests an approach which first reduces the exterior boundary value problem to an interior boundary value problem in the disk D_R of radius R with a nonlocal boundary condition on its boundary Γ_R, and then uses a finite element method to approximate the field u on D_R.

As a related, but different, method for each function φ_j from a (finite) set of linearly independent functions on Γ_R one can solve the boundary value problem in the region between $\partial\Omega$ and Γ_R with boundary data g and φ_j by a standard finite element method. Then one solves the boundary value problem outside the circle with boundary data φ_j. If we choose for φ_j trigonometric polynomials this can be done explicitly. An appropriate linear combination will guarantee that both of the Cauchy data (almost) coincide on Γ_R and, therefore, gives approximation of the exact solution. This method is particularly suitable for non-homogeneous media.

In a more general approach, the finite element method and a boundary element method can be combined. We refer to [70, 58, 91, 92, 44] for various modifications of this idea.

For further reading we refer to the monograph [54] which is solely devoted to Finite Element methods for the Helmholtz equation.

5.5.1 Nyström's Method for Periodic Weakly Singular Kernels

We have introduced the Nyström Method earlier in Subsection 4.5.2 and applied it to the linear line source in Subsection 4.5.4. We now wish to indicate the application of the method to quadrature problems in which the integral contains a weakly singular kernel. Our presentation will follow the one in the monograph of Colton and Kress [30].

In the previous application, the kernels of the integral equations were smooth – even analytic in the case of the linear line source. In the boundary integral equation methods where one solves the integral equations (5.22) and (5.28) numerically the kernels have a weak singularity of logarithmic type. For the following we restrict ourselves to the Dirichlet boundary value problem and a numerical solution of equation (5.22).

We assume that the boundary $\partial\Omega$ possesses a regular analytic and periodic (with period 2π) parametric representation of the form

$$\boldsymbol{x} \;=\; \boldsymbol{x}(t) \;=\; \big(x_1(t)\,,\,x_2(t)\big)\,, \quad 0 \le t \le 2\pi\,, \tag{5.40}$$

satisfying $|\dot{\boldsymbol{x}}(t)| := \sqrt{\dot{x}_1(t)^2 + \dot{x}_2(t)^2} > 0$ on $[0, 2\pi]$. Then, using $\frac{d}{dz}H_0^{(1)}(z) = -H_1^{(1)}(z)$, we transform (5.21) into the parametric form

$$\psi(t) \;-\; \int\limits_0^{2\pi} \Big\{ L(t,s) + ik\,M(t,s) \Big\}\,\psi(s)\,ds \;=\; \tilde{g}(t)\,, \quad 0 \le t \le 2\pi\,, \tag{5.41}$$

where we have set $\psi(t) = \varphi(\boldsymbol{x}(t))$, $\tilde{g}(t) = 2g(\boldsymbol{x}(t))$,

$$\begin{aligned} L(t,s) &:= \frac{ik}{2} \Big\{ \dot{x}_2(s)\big[x_1(s) - x_1(t)\big] - \dot{x}_1(s)\big[x_2(s) - x_2(t)\big] \Big\} \cdot \\ &\quad \cdot \frac{H_1^{(1)}\big(k\,|\boldsymbol{x}(t) - \boldsymbol{x}(s)|\big)}{|\boldsymbol{x}(t) - \boldsymbol{x}(s)|}\,, \\ M(t,s) &:= \frac{i}{2}\,H_0^{(1)}\big(k\,|\boldsymbol{x}(t) - \boldsymbol{x}(s)|\big)\ |\dot{\boldsymbol{x}}(t)| \end{aligned}$$

for $t \neq s$. From the behaviour of the Neumann function near the origin we see that the kernels L and M have logarithmic singularities at $t = s$. Hence, for an optimal numerical treatment, we follow Martensen [93] and Kussmaul [78] and split the kernels into two parts

$$\begin{aligned} L(t,s) &= L_1(t,s)\,\ln\left(4\,\sin^2\frac{t-s}{2}\right) \;+\; L_2(t,s), \\ M(t,s) &= M_1(t,s)\,\ln\left(4\,\sin^2\frac{t-s}{2}\right) \;+\; M_2(t,s), \end{aligned}$$

where

$$
\begin{aligned}
L_1(t,s) &:= \frac{k}{2\pi}\left\{\dot{x}_2(s)\big[x_1(s)-x_1(t)\big] - \dot{x}_1(s)\big[x_2(s)-x_2(t)\big]\right\}\cdot \\
&\quad\cdot\frac{J_1\big(k\,|\boldsymbol{x}(t)-\boldsymbol{x}(s)|\big)}{|\boldsymbol{x}(t)-\boldsymbol{x}(s)|}\,, \\
L_2(t,s) &:= L(t,s) \;-\; L_1(t,s)\,\ln\left(4\,\sin^2\frac{t-s}{2}\right), \\
M_1(t,s) &:= -\frac{1}{2\pi}\,J_0\big(k\,|\boldsymbol{x}(t)-\boldsymbol{x}(s)|\big)\;|\dot{\boldsymbol{x}}(t)|\,, \\
M_2(t,s) &:= M(t,s) \;-\; M_1(t,s)\,\ln\left(4\,\sin^2\frac{t-s}{2}\right).
\end{aligned}
$$

The kernels L_1, L_2, M_1, M_2 turn out to be analytic with values on the diagonal:

$$
\begin{aligned}
L_1(t,t) = 0\,, \quad L_2(t,t) \;&=\; L(t,t) \;=\; \frac{1}{2\pi}\,\frac{\dot{x}_1(t)\,\ddot{x}_2(t) - \dot{x}_2(t)\,\ddot{x}_1(t)}{\dot{x}_1(t)^2 + \dot{x}_2(t)^2}\,, \\
M_1(t,t) = -\frac{1}{2\pi}\,|\dot{\boldsymbol{x}}(t)|\,, \quad M_2(t,t) \;&=\; \left\{\frac{i}{2} - \frac{\gamma}{\pi} - \frac{1}{2\pi}\,\ln\left(\frac{k^2}{4}\,|\dot{\boldsymbol{x}}(t)|^2\right)\right\}|\dot{\boldsymbol{x}}(t)|
\end{aligned}
$$

for $0 \le t \le 2\pi$. Hence, we must solve an integral equation of the form

$$
\psi(t) \;-\; \int_0^{2\pi} K_1(t,s)\,\ln\left(4\,\sin^2\frac{t-s}{2}\right)\,\psi(s)\,ds \;+\; \int_0^{2\pi} K_2(t,s)\,\psi(s)\,ds \;=\; \tilde{g}(t)\,, \tag{5.42}
$$

for $0 \le t \le 2\pi$ with $K_j = L_j + ikM_j$, $j = 1,2$. As we have seen in Subsection 4.5.2, the Nyström method replaces the integral by an appropriately chosen quadrature formula. Since the integrand of the second integral is analytic and 2π–periodic, the ordinary trapezoidal rule with equi-distant grid points $t_j^{(N)} := \pi j/N$, for $j = 0,\dots,2N-1$, is optimal. This yields an approximation of the form

$$
\int_0^{2\pi} f(s)\,ds \;\approx\; \frac{\pi}{N}\sum_{j=0}^{2N-1} f\big(t_j^{(N)}\big)\,. \tag{5.43}
$$

One constructs the analogous formula for the first integral by replacing the smooth factor of the integrand by its trigonometric interpolation. This yields

$$
\int_0^{2\pi} \ln\left(4\,\sin^2\frac{t-s}{2}\right)\,f(s)\,ds \;\approx\; \sum_{j=0}^{2N-1} R_j^{(N)}(t)\,f\big(t_j^{(N)}\big) \tag{5.44}
$$

with

$$
R_j^{(N)}(t) \;:=\; -\frac{2\pi}{N}\sum_{m=1}^{N-1}\frac{1}{m}\,\cos m\big(t - t_j^{(N)}\big) \;-\; \frac{\pi}{N^2}\,\cos N\big(t - t_j^{(N)}\big)\,,
$$

$j = 0, \dots, 2N-1$. In the Nyström method the integral equation (5.42) is then replaced by the equation

$$\psi^{(N)}(t) \;-\; \sum_{j=0}^{2N-1} \Bigl\{ R_j^{(N)}(t)\, K_1\bigl(t, t_j^{(N)}\bigr) + \frac{\pi}{N}\, K_2\bigl(t, t_j^{(N)}\bigr) \Bigr\}\, \psi\bigl(t_j^{(N)}\bigr) \;=\; \tilde{g}(t) \tag{5.45}$$

for $0 \le t \le 2\pi$. This equation reduces to a finite dimensional system if we substitute $t = t_i^{(N)}$, $i = 0, \dots, 2N-1$. With $\psi_i := \psi\bigl(t_i^{(N)}\bigr)$ this finite dimensional approximation is

$$\psi_i - \sum_{j=0}^{2N-1} \Bigl\{ R_{|i-j|}\, K_1\bigl(t_i^{(N)}, t_j^{(N)}\bigr) + \frac{\pi}{N}\, K_2\bigl(t_i^{(N)}, t_j^{(N)}\bigr) \Bigr\}\, \psi_j \;=\; \tilde{g}\bigl(t_i^{(N)}\bigr), \tag{5.46}$$

for $i = 0, \dots, 2N-1$, where

$$R_\ell \;:=\; R_\ell(0)^{(N)} \;=\; -\frac{2\pi}{N} \sum_{m=1}^{N-1} \frac{1}{m} \cos\frac{m\ell\pi}{N} \;-\; \frac{(-1)^\ell \pi}{N^2},$$

for $\ell = 0, \dots, 2N-1$.

If, now, ψ_i, $i = 0, \dots, 2N-1$, is a solution of (5.46) and one defines the function $\psi^{(N)}$ by

$$\psi^{(N)}(t) \;:=\; \sum_{j=0}^{2N-1} \Bigl\{ R_j^{(N)}(t)\, K_1\bigl(t, t_j^{(N)}\bigr) + \frac{\pi}{N}\, K_2\bigl(t, t_j^{(N)}\bigr) \Bigr\}\, \psi_j \;+\; \tilde{g}(t), \tag{5.47}$$

$0 \le t \le 2\pi$, then $\psi^{(N)}$ solves (5.45). Therefore, this last formula provides a natural interpolation of the values ψ_j, $j = 0, \dots, 2N-1$. For the solution of the large system (5.46) we recommend the use of fast iterative methods such as multi-grid solvers, cf. [76].

To show that this method is, in fact, one which yields approximations to the actual solution, we must prove that $\psi^{(N)} \to \psi$ as $N \to \infty$ in a suitable norm, and preferably do so in a way that the rate of convergence is established. Indeed, the following convergence theorem can be proven (cf. [76]):

Theorem 5.21. *Let $\Omega \subset \mathbb{R}^2$ be a bounded region with connected exterior and with $\partial\Omega$ be analytic.*

(a) Let $g \in C^m(\partial\Omega)$ for some $m \in \mathbb{N}$ i.e., $\tilde{g} \in C^m_{per}[0, 2\pi]$ where $C^m_{per}[0, 2\pi]$ denotes the space of $2\pi-$periodic m times continuously differentiable functions. Then equation (5.46) is uniquely solvable for all sufficiently large N and there exists $c > 0$ independent of N and g with

$$\|\psi^{(N)} - \psi\|_{C(\partial\Omega)} \;\le\; \frac{c}{N^m}\, \|\tilde{g}^{(m)}\|_{C(\partial\Omega)}\,. \tag{5.48a}$$

(b) If g is analytic on $\partial\Omega$, then there exists $c > 0$ and $\sigma > 0$, both independent of N, with

$$\|\psi^{(N)} - \psi\|_{C(\partial\Omega)} \;\le\; c\,\mathrm{e}^{-\sigma N}. \tag{5.48b}$$

Having computed the density ψ from (5.42) we can calculate the far field pattern u_∞ by formula (5.24). Using the parametrization $\boldsymbol{x} = \boldsymbol{x}(t)$, $0 \le t \le 2\pi$, again we have:

$$u_\infty(t) \;=\; \int_0^{2\pi} U(t,s)\,\psi(s)\,ds\,, \quad 0 \le t \le 2\pi\,, \tag{5.49}$$

where

$$U(t,s) = -ik\sigma\left[\left(\dot{x}_2(s)\,\cos t - \dot{x}_1(s)\,\sin t\right) \;+\; |\dot{\boldsymbol{x}}(s)|\right]\mathrm{e}^{-ik(x_1(s)\cos t + x_2(s)\sin t)}$$

for $t, s \in [0, 2\pi]$. The error estimates from Theorem 5.21 carry over to estimates in the fields $(u^{(N)} - u)|_K$ and to the far field patterns $u_\infty^{(N)} - u_\infty$.

As we see from this result this Nyström method is particularly fast for smooth data. Nevertheless, the Nyström method can be modified to treat also boundary value problems in domains with corners. We refer to [75] for more details.

5.5.2 Complete Families of Solutions

Now we turn back to the Helmholtz equation (5.1a) in an exterior domain in $\mathbb{R}^2$:

$$\Delta u \;+\; k^2 u \;=\; 0 \quad \text{in } \Omega^c\,. \tag{5.50}$$

The method for solving boundary value problems which we will describe in this subsection is based on the construction of families of special radiating solutions of (5.50) with the property that any radiating solution of the Helmholtz equation can be approximated arbitrarily well by *finite* linear combinations of this family. At the moment, we do not give a precise meaning of this approximation property but will call such a family of solutions a **complete family**. The advantage of such an approach is obvious: Every linear combination satisfies the Helmholtz equation exactly. Moreover, the corresponding far field pattern of the approximation is given explicitly by the known far field patterns of the members of this family. The coefficients of the combination have to be determined by the boundary condition which can only be satisfied approximately with a finite linear combination.

The general idea of the method seems to have originated with Collatz [25, 26], (see also the book of Miranda [99]) and even earlier work of Rayleigh [131]. For the Helmholtz equation, specific work has been done by Müller and Kersten [106], Limić [82, 83], and Kersten [64]. These families have been used in some inverse shape identification problems by Angell and Kleinman [8], Angell, Kleinman and Roach [11], and Angell, Jiang, and Kleinman [4]. Even earlier, for the electromagnetic case, the ideas were used systematically by Calderón [21] and later by Müller [104] and in the thesis of Fast [40]. For parabolic equations the method was used by Collatz and for the Stefan and inverse Stefan problem by Lozano and Reemtsen [87] and by Reemtsen and Kirsch [112]. We will come back to applications of this method to optimization problems in Chapter 7.

We make the general assumption that the boundary of the domain Ω, $\partial\Omega$, is smooth enough to apply the results of Section 5.3. In particular, we assume and that the boundary value problems are uniquely solvable for every g.

We begin by introducing two families of radiating solutions which seem to be particularly appropriate for the Helmholtz equation. Define the following sets $\mathcal{C}_1$ and $\mathcal{C}_2$ of functions by

$$\mathcal{C}_1 := \left\{ H_m^{(1)}(kr)\, \mathrm{e}^{im\theta} : m \in \mathbb{Z} \right\}, \tag{5.51a}$$

$$\mathcal{C}_2 := \left\{ \Phi(\boldsymbol{x}, \boldsymbol{x}_m) : m \in \mathbb{Z} \right\}. \tag{5.51b}$$

Here, $H_m^{(1)}$ denotes the Hankel function of first kind and order m and $\Phi(\boldsymbol{x},\boldsymbol{y}) = \frac{i}{4} H_0^{(1)}(k|\boldsymbol{x}-\boldsymbol{y}|)$ the fundamental solution of the Helmholtz equation. The "ficticious" source points $\boldsymbol{x}_m \in \Omega$ are assumed to be distinct and lie in the interior Ω. Also, we assume that the origin is contained in Ω.

We first address the question of linear independence of these functions restricted to the boundary $\partial\Omega$:

Theorem 5.22. *Any finite subset $\{v_m : |m| \leq N\} \subset \mathcal{C}$ where $\mathcal{C} = \mathcal{C}_1$ or $\mathcal{C} = \mathcal{C}_2$ is linearly independent on $\partial\Omega$ i.e., $\sum_{|m|\leq N} a_m\, v_m(x) = 0$ for $x \in \partial\Omega$ implies that $a_m = 0$ for all m.*

Proof: Assume that the solution v of the Helmholtz equation, defined by

$$v(\boldsymbol{x}) \;:=\; \sum_{m=-N}^{N} a_m\, v_m(\boldsymbol{x})\,, \quad \boldsymbol{x} \in \mathbb{R}^2 \setminus \partial\Omega\,,$$

vanishes on $\partial\Omega$. Since v is a radiating solution, v vanishes in the exterior $\mathbb{R}^2 \setminus \Omega$ according to the uniqueness theorem (Theorem 5.8).

If $v_m(\boldsymbol{x}) = H_m^{(1)}(kr)\exp(im\theta)$ then $a_m = 0$ since, for sufficiently large R, $0 = v(R,\theta) = \sum_{m=-N}^{N} a_m\, H_m^{(1)}(kR)\exp(im\theta)$ is a Fourier sum and $H_m^{(1)}(kR) \neq 0$ for any m. In the case $v_m(\boldsymbol{x}) = H_0^{(1)}(k|\boldsymbol{x}-\boldsymbol{x}_m|)$ we can analytically continue the function $v(\boldsymbol{x}) = \frac{i}{4}\sum_{m=-N}^{N} a_m\, H_0^{(1)}(k|\boldsymbol{x}-\boldsymbol{x}_m|)$ to $\mathbb{R}^2 \setminus \{\boldsymbol{x}_m : |m| \leq N\}$. Since v vanishes identically, the limit $\boldsymbol{x} \to \boldsymbol{x}_\ell$ yields $a_\ell = 0$. Therefore, in this case, as well, $a_\ell = 0$ for all ℓ. □

We make the following further **assumption** on the set $\{\boldsymbol{x}_m : m \in \mathbb{Z}\} \subset \Omega$ of source points:

> *For any solution u of the Helmholtz equation in the interior domain Ω, the vanishing $u(\boldsymbol{x}_m) = 0$ for all $m \in \mathbb{Z}$ implies that u vanishes in all of Ω.*

Before proving completeness of the two families $\mathcal{C}_1$ and $\mathcal{C}_2$, we check that sets of points $\boldsymbol{x}_m$ with this property in fact exist.

Lemma 5.23. *Let $\hat{\Gamma} \subset \Omega$ be a closed analytic curve whose interior, $\hat{\Omega}$ is completely contained in Ω i.e., the closure of $\hat{\Omega}$ is contained in Ω. Then, if k^2 is not a Dirichlet eigenvalue for $\hat{\Omega}$, any countable set of infinitely many points $\boldsymbol{x}_m \in \hat{\Gamma}$ has the required property i.e., for any solution u of the Helmholtz equation in Ω, $u(\boldsymbol{x}_m) = 0$ for all m implies that u vanishes in all of Ω.*

Proof: Let u be a solution of $\Delta u + k^2 u = 0$ in Ω such that $u(\boldsymbol{x}_m) = 0$ for all m. Since $\hat{\Gamma}$ is analytic, u must vanish on $\hat{\Gamma}$ by the identity theorem for analytic functions. Since k^2 is not a Dirichlet eigenvalue u has to vanish in the interior, $\hat{\Omega}$, of $\hat{\Gamma}$. Then also $u = 0$ in Ω by analytic continuation. □

This assumption on the sources $\boldsymbol{x}_m$ is needed for the next theorem which states that certain traces of functions in $\mathcal{C}_j$ are complete in $L^2(\partial\Omega)$, i.e. the set of all finite linear combinations is dense.

Theorem 5.24. *Let $\alpha \in \mathbb{R}$, $\lambda \in L^\infty(\partial\Omega)$ with $\alpha \cdot \operatorname{Im} \lambda \geq 0$ almost everywhere on $\partial\Omega$ and $|\alpha| + |\lambda(\boldsymbol{x})| > 0$ almost everywhere on $\partial\Omega$. The sets*

$$\mathcal{B}_j \;:=\; \left\{ \alpha \frac{\partial v}{\partial n} + \lambda\, v|_{\partial\Omega} : v \in \operatorname{span} \mathcal{C}_j \right\}$$

are dense in $L^2(\partial\Omega)$ for $j = 1$ and $j = 2$.

Proof: First, we discuss the case of $\mathcal{B}_2$. We use the fact that a subspace is dense in $L^2(\partial\Omega)$ if, and only if, its orthogonal complement is $\{0\}$, see Definition A.10. Therefore, let $g \in L^2(\partial\Omega)$ be such that $\overline{g}$ is orthogonal to $\mathcal{B}_2$, i.e.

$$\int\limits_{\partial\Omega} g(\boldsymbol{y}) \left[\alpha \frac{\partial \Phi(\boldsymbol{y}, \boldsymbol{x}_m)}{\partial n(\boldsymbol{y})} + \lambda \Phi(\boldsymbol{y}, \boldsymbol{x}_m) \right] d\ell(\boldsymbol{y}) \;=\; 0 \quad \text{for all } m \in \mathbb{Z}\,. \tag{5.52}$$

We define the function u by

$$u(\boldsymbol{x}) \;:=\; \int_{\partial\Omega} g(\boldsymbol{y}) \left[\alpha \frac{\partial \Phi(\boldsymbol{x}, \boldsymbol{y})}{\partial n(\boldsymbol{y})} + \lambda \Phi(\boldsymbol{x}, \boldsymbol{y}) \right] d\ell(\boldsymbol{y}) \quad \text{for } \boldsymbol{x} \notin \partial\Omega\,. \tag{5.53}$$

Then u is a solution of the Helmholtz equation and $u(\boldsymbol{x}_m) = 0$ for all m. Therefore $u(\boldsymbol{x}) = 0$ for all $\boldsymbol{x} \in \Omega$ by the assumption on the sources $\boldsymbol{x}_m$. Using the L^2–jump conditions (5.33a)–(5.33d) we conclude that

$$\lim_{t\to 0+} \|u_{+t} - u_{-t} \; - \; \alpha\, g\|_{L^2(\partial\Omega)} = 0\,, \tag{5.54a}$$

$$\lim_{t\to 0+} \|\boldsymbol{n}\cdot\nabla u_{+t} \; - \; \boldsymbol{n}\cdot\nabla u_{-t} \; + \; \lambda g\|_{L^2(\partial\Omega)} = 0\,. \tag{5.54b}$$

Since $u = 0$ in Ω we have the estimate

$$\begin{aligned}&\|\lambda\, u_t + \alpha\, \boldsymbol{n}\cdot\nabla u_t\|_{L^2(\partial\Omega)}\\ &= \|\lambda(u_{+t} - u_{-t} - \alpha\, g) + \alpha(\boldsymbol{n}\cdot\nabla u_{+t} - \boldsymbol{n}\cdot\nabla u_{-t} + \lambda\, g)\|_{L^2(\partial\Omega)}\\ &\le \|\lambda\|_{L^\infty(\partial\Omega)}\, \|u_{+t} - u_{-t} - \alpha\, g\|_{L^2(\partial\Omega)} \; + \; |\alpha|\; \|\boldsymbol{n}\cdot\nabla u_{+t} - \boldsymbol{n}\cdot\nabla u_{-t} + \lambda\, g\|_{L^2(\partial\Omega)}\,.\end{aligned}$$

Hence, using the jump relations (5.54a), (5.54b) we see that

$$\lim_{t\to 0+} \|\lambda\, u_t + \alpha\, \boldsymbol{n}\cdot\nabla u_t\|_{L^2(\partial\Omega)} \;=\; 0\,.$$

Therefore, u satisfies a homogeneous exterior impedance problem with boundary condition $\partial u/\partial n + (\lambda/\alpha)\, u = 0$ on $\partial\Omega$ (or a Dirichlet problem if $\alpha = 0$). The uniqueness theorem (Theorem 5.19) also shows that $u = 0$ in Ω^c. It follows from the jump relations (5.54a), (5.54b) that $g = 0$. This proves the denseness of $\mathcal{B}_2$.

Let now $g \in L^2(\partial\Omega)$ such that $\overline{g}$ is orthogonal to $\mathcal{B}_1$, i.e.

$$\int_{\partial\Omega} g(\boldsymbol{y}) \left[\alpha\, \frac{\partial H_m^{(1)}(\boldsymbol{y})}{\partial n} + \lambda\, H_m^{(1)}(\boldsymbol{y})\right] d\ell(\boldsymbol{y}) \;=\; 0 \quad \text{for all } m \in \mathbb{Z}\,. \tag{5.55}$$

We define u by equation (5.53) again. The addition formula (cf. [1])

$$H_0^{(1)}(k\,|\boldsymbol{x} - \boldsymbol{y}|) \;=\; \sum_{m=-\infty}^{\infty} J_m(k\,|\boldsymbol{x}|)\, H_m^{(1)}(k\,|\boldsymbol{y}|)\, \mathrm{e}^{im(\theta-\psi)}, \quad |\boldsymbol{x}| < |\boldsymbol{y}|\,, \tag{5.56}$$

where θ and ψ are the arguments of $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively, then shows that u vanishes in a small disc with center 0 contained in Ω. By unique continuation u vanishes in all of Ω. Now we proceed as in the first case above. This ends the proof. □

Remarks: Many extensions of these completeness results are available. In particular we refer the reader to the following particular extensions.

(a) These denseness results could be proved in Sobolev spaces by the same methods (cf. Limić [82, 83] who proved them for the Neumann and Dirichlet problems in $H^{-1/2}(\partial\Omega)$ and $H^{1/2}(\partial\Omega)$, respectively). In the proof one replaces the inner product $\int_{\partial\Omega} g\,\overline{h}\, d\ell$ by the dual bracket $\langle g, h\rangle_{1/2}$ for $g \in H^{-1/2}(\partial\Omega)$ and $h \in H^{1/2}(\partial\Omega)$.

(b) Kersten [63, 64] has even proven the completeness of $\mathcal{B}_j$ in $C(\partial\Omega)$ and for the case that $\partial\Omega$ admits corners. For this case one has to investigate potentials with densities in the dual space of $C(\partial\Omega)$, i.e. for potentials of the form $\int_{\partial\Omega} \Phi(\boldsymbol{x}, \boldsymbol{y})\, d\mu(\boldsymbol{y})$ with Borel measures μ on $\partial\Omega$.

(c) For interior regions Ω and the special family

$$\mathcal{B}_N = \operatorname{span}\{J_m^{(1)}(kr)\, e^{im\theta} : |m| \le N\},$$

Still [129] has proven error estimates of the approximation error. In particular, if $\partial\Omega$ is smooth enough and $h \in C^{p,\alpha}(\partial\Omega)$ for some $p \in \mathbb{N}$ and $\alpha \in (0,1]$, then there exists $c > 0$ with

$$\inf\{\|v_N - u\|_{C(\overline{\Omega})} : v_N \in \mathcal{B}_N\} \le \frac{c}{N^{p+\alpha}}.$$

Now we show that the sets $\mathcal{C}_1$ and $\mathcal{C}_2$, given by (5.51a) and (5.51b), themselves are complete in the following sense.

Theorem 5.25. *Let u be any radiating solution of the Helmholtz equation (5.50) in Ω^c and K be any compact set in Ω^c. Then there is a sequence $\{u_n\} \subset \operatorname{span}\mathcal{C}$, where $\mathcal{C} = \mathcal{C}_1$ or $\mathcal{C} = \mathcal{C}_2$, which converges uniformly to u in K. Furthermore, $\{u_n\}$ can be chosen so that the far field patterns of u_n converge to the far field pattern of u, uniformly on the unit circle.*

Proof: Without loss of generality we assume that u is continuous up to the boundary $\partial\Omega$. (Otherwise choose Ω' with C^2–boundary $\partial\Omega'$ such that $\Omega \subset \overline{\Omega'}$ and K is contained in the exterior of Ω'. Then u is continuous up to the boundary $\partial\Omega'$.) By Theorem 5.19 we know that the solution $u|_K$ and the far field pattern u_∞ of the exterior Dirichlet problem depend continuously on the boundary data $u|_{\partial\Omega}$ with respect to the L^2–norm. This means that there exists a constant $c > 0$ with

$$\|v|_K\|_{C(K)} + \|v_\infty\|_{C(S^1)} \le c\,\|v|_{\partial\Omega}\|_{L^2(\partial\Omega)} \tag{5.57}$$

for all radiating solutions v. By Lemma 5.24 with $\alpha = 0$ and $\lambda \equiv 1$, there exists a sequence $\{u_n\} \subset \operatorname{span}\mathcal{C}$ with $\|(u_n - u)|_{\partial\Omega}\|_{L^2(\partial\Omega)} \to 0$ as n tends to infinity. With $v = u_n - u$ in (5.57) the conclusion of the theorem follows. □

This last theorem suggests a method for the numerical calculation of the solution to the boundary value problem (5.2). In order to give the idea most clearly we restrict ourselves to the Dirichlet case (5.2).

Specifically, since finite sums of the form

$$\sum_{m=-N}^{N} a_m\, H_m^{(1)}(kr)\, e^{im\theta} \quad \text{and} \quad \sum_{m=-N}^{N} a_m\, \Phi(\boldsymbol{x}, \boldsymbol{x}_m) \tag{5.58}$$

are radiating solutions of the Helmholtz equation we approximate a solution of the boundary value problem by determining the unknown coefficients, a_m, so that the error in the boundary condition is minimized. Thus, if $\|\cdot\|$ is any norm on $C(\partial\Omega)$ we pose the following minimization problem:

$$\text{Minimize} \quad \left\| \sum_{m=-N}^{N} a_m\, v_m - g \right\| \quad \text{subject to} \quad a_m \in \mathbb{C},\ |m| \le N. \tag{5.59}$$

Here, either $v_m(\boldsymbol{x}) = H_m^{(1)}(kr)\, \mathrm{e}^{im\theta}$ or $v_m(\boldsymbol{x}) = \Phi(\boldsymbol{x}, \boldsymbol{x}_m)$ restricted to the boundary $\partial\Omega$. We can now prove the following theorem:

Theorem 5.26.

(a) For every $N \in \mathbb{N}$ there exists a solution $\boldsymbol{a}^{(N)} \in \mathbb{C}^{2N+1}$ of (5.59).
(b) If $\|\cdot\| = \|\cdot\|_{L^2(\partial\Omega)}$ then the solution is unique.
(c) Let u be the (weak) solution of the boundary value problem (5.2) and

$$u_N = \sum_{m=-N}^{N} a_m^{(N)} v_m$$

where $\boldsymbol{a}^{(N)} \in \mathbb{C}^{2N+1}$ is the solution of (5.59) for $\|\cdot\| = \|\cdot\|_{L^2(\partial\Omega)}$. Then $u_N \to u$ as $N \to \infty$ uniformly on every compact subset $K \subset \Omega^c$. Furthermore, the corresponding far field patterns $u_{N,\infty}$ also converge to u_∞ uniformly on the unit circle S^1, and there exists a constant $c > 0$ with

$$\begin{aligned} &\|(u_N - u)|_K\|_{C(K)} + \|u_{N,\infty} - u_\infty\|_{C(\partial\Omega)} \\ &\qquad \le c \inf\big\{ \|v_N|_{\partial\Omega} - g\|_{L^2(\partial\Omega)} : v_N \in \mathcal{C} \big\}. \end{aligned}$$

Proof: For $N \in \mathbb{N}$ we define the operator $L_N : \mathbb{C}^{2N+1} \longrightarrow C(\partial\Omega)$ by

$$(L_N \boldsymbol{a})(\boldsymbol{x}) := \sum_{n=-N}^{N} a_m\, v_m(\boldsymbol{x}), \quad \boldsymbol{x} \in \partial\Omega.$$

The problem (5.59) is then written as the problem of minimizing $\|L_N\boldsymbol{a} - g\|$ on $\mathbb{C}^{2N+1}$ or, if we define the range space $\mathcal{R}(L_N)$ of L_N by $\mathcal{R}(L_N) = \{w = L_N\boldsymbol{a} : \boldsymbol{a} \in \mathbb{C}^{2N+1}\}$, as that of minimizing $\|w - g\|$ subject to $w \in \mathcal{R}(L_N)$. First we note that L_N is one-to-one, which is just a reformulation of the linear independence of the functions $v_m|_{\partial\Omega}$ which was established in Theorem 5.22. Second, we note that $\mathcal{R}(L_N)$ is finite dimensional (with dimension $2N+1$) and is therefore closed. We define

$$J_N := \inf_{\boldsymbol{a} \in \mathbb{C}^{2N+1}} \|L_N\boldsymbol{a} - g\| = \inf_{w \in \mathcal{R}(L_N)} \|w - g\|.$$

(a) Let $\{w_m\} \subset \mathcal{R}(L_N)$ be a minimizing sequence, i.e. a sequence with the property that $\|w_m - g\| \to J_N$ as m tends to infinity. In particular, the sequence $\{w_m\}$ is bounded. Since $\mathcal{R}(L_N)$ is finite dimensional, it contains a convergent subsequence by the theorem of Bolzano-Weierstrass: $w_{m_j} \to \hat{w}$ as $j \to \infty$. Then $\hat{w} \in \mathcal{R}(L_N)$ and, by the continuity of the norm, $\|\hat{w} - g\| = J_N$.

(b) (Uniqueness of minima) Let $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{C}^{2N+1}$ both be minima. Set $\boldsymbol{c} := (\boldsymbol{a}+\boldsymbol{b})/2$. Then, by the parallelogram equality

$$\|f+h\|^2_{L^2(\partial\Omega)} + \|f-h\|^2_{L^2(\partial\Omega)} = 2\|f\|^2_{L^2(\partial\Omega)} + 2\|h\|^2_{L^2(\partial\Omega)}, \tag{5.60}$$

we conclude that

$$\begin{aligned} J_N \leq \|L_N\boldsymbol{c} - g\|^2_{L^2(\partial\Omega)} &= \left\|\frac{1}{2}(L_N\boldsymbol{a} - g) + \frac{1}{2}(L_N\boldsymbol{b} - g)\right\|^2_{L^2(\partial\Omega)} \\ &= \frac{1}{2}\|L_N\boldsymbol{a} - g\|^2_{L^2(\partial\Omega)} + \frac{1}{2}\|L_N\boldsymbol{b} - g\|^2_{L^2(\partial\Omega)} - \frac{1}{4}\|L_N(\boldsymbol{a}-\boldsymbol{b})\|^2_{L^2(\partial\Omega)} \\ &= J_N - \frac{1}{4}\|L_N(\boldsymbol{a}-\boldsymbol{b})\|^2_{L^2(\partial\Omega)}. \end{aligned}$$

This yields $L_N(\boldsymbol{a}-\boldsymbol{b}) = 0$, thus $\boldsymbol{a} = \boldsymbol{b}$ by the injectivity of L_N.

(c) The continuous dependence result of Theorem 5.19 yields the existence of $c > 0$ with

$$\|(u_N - u)|_K\|_{C(K)} + \|u_{N,\infty} - u_\infty\|_{C(S^1)} \leq c\,\|(u_N - u)|_{\partial\Omega}\|_{L^2(\partial\Omega)} = c\,J_N$$

for all N since $u|_{\partial\Omega} = g$. It remains to show that J_N tends to zero if N tends to infinity. Let $\epsilon > 0$. By Theorem 5.24 we can find $N \in \mathbb{N}$ and $v_N \in \mathcal{R}(L_N)$, with $J_N \leq \|v_N|_{\partial\Omega} - g\|_{L^2(\partial\Omega)} \leq \epsilon$. This proves part (c) and completes the proof of the theorem. □

One is tempted to apply the general results of Subsections 3.2.1 and 3.2.5 on existence and approximation. In our case, however, we study an *unconstrained* minimization problem i.e., one of the main assumptions of those theorems is not satisfied. Indeed, in general, the coefficients $a_m^{(N)}$ of the optimal solution u_N do not converge unless the **Rayleigh hypothesis** is satisfied. It is well known that the solution of the three-dimensional exterior Helmholtz equation can be written as a convergent series of spherical harmonics outside the smallest sphere containing the set Ω. The Raleigh hypothesis in this case is simply that this expansion is valid up to and on the boundary $\partial\Omega$. For some situations this is false, in others, true. We refer to [97] for further details.

5.5.3 Finite Element Methods for Absorbing Boundary Conditions

For boundary value problems in bounded domains, the finite element method is probably the best known numerical method of all. Its advantages lie in its broad range of applicability, the low regularity requirements on the data, and the sparseness of the resulting finite system.

Before we describe its application to the exterior Dirichlet boundary value problem we first recall the idea of the finite element method for the simplest case of the Dirichlet boundary value problem in a bounded domain, namely the determination of a solution u in some bounded domain $D \subset \mathbb{R}^2$ with

$$\Delta u + k^2 u = 0 \text{ in } D\,, \qquad \text{and } u = g \text{ on } \partial D.. \tag{5.61}$$

The finite element method is based on the variational formulation of the problem which we describe first. Assuming that u is a solution of (5.27), multiplying the Helmholtz equation by any function $\phi \in C^1(\overline{D})$ with $\phi|_{\partial D} = 0$ and integrating over D we have, trivially, that

$$0 \;=\; \iint\limits_D \left(\Delta u + k^2 u\right) \phi \, dx.$$

Green's first theorem yields

$$\iint\limits_D \left(\nabla u \cdot \nabla \phi - k^2 u\, \phi\right) dx \;=\; 0\,, \tag{5.62}$$

where we have used the assumption that ϕ vanishes on ∂D. Next, we choose an arbitrary extension $\tilde{g}$ of g to D and write u in the form $u = v + \tilde{g}$ where now $v|_{\partial D} = 0$. Substituting this form of u into (5.62) and replacing ϕ by its complex conjugate yields

$$\iint\limits_D \left(\nabla v \cdot \nabla \overline{\phi} - k^2 v\, \overline{\phi}\right) dx \;=\; -\iint\limits_D \left(\nabla \tilde{g} \cdot \nabla \overline{\phi} - k^2 \tilde{g}\, \overline{\phi}\right) dx$$

or, adding and subtracting $v\overline{\phi}$,

$$\iint\limits_D \left(\nabla v \cdot \nabla \overline{\phi} + v\, \overline{\phi}\right) dx \;-\; (k^2 + 1) \iint\limits_D v\, \overline{\phi}\, dx \;=\; -\iint\limits_D \left(\nabla \tilde{g} \cdot \nabla \overline{\phi} - k^2 \tilde{g}\, \overline{\phi}\right) dx \tag{5.63}$$

for all functions $\phi \in X$ where

$$X \;:=\; \left\{\phi \in C^1(\overline{D}) : \phi|_{\partial D} = 0\right\}. \tag{5.64}$$

Clearly the first integral in (5.63) defines an inner product on X with corresponding norm

$$\|\phi\|_X \;=\; \sqrt{\iint\limits_D \left(\left|\nabla \phi\right|^2 + |\phi|^2\right) dx}\,. \tag{5.65}$$

The right hand side of (5.63) defines a bounded **anti-linear functional**[5]

$$\ell(\phi) \;:=\; -\iint\limits_D \left(\nabla \tilde{g} \cdot \nabla \overline{\phi} - k^2 \tilde{g}\, \overline{\phi}\right) dx\,, \quad \phi \in X\,.$$

[5] A functional $\ell : X \longrightarrow \mathbb{C}$ is called anti-linear if $\phi \mapsto \overline{\ell(\phi)}$ is linear

Indeed, boundedness follows from the Cauchy-Schwarz inequality,

$$\begin{aligned} |\ell(\phi)| &\leq \|\nabla\phi\|_{L^2(D)} \, \|\nabla\tilde{g}\|_{L^2(D)} \; + \; k^2 \, \|\phi\|_{L^2(D)} \, \|\tilde{g}\|_{L^2(D)} \\ &\leq \|\phi\|_X \left[\|\nabla\tilde{g}\|_{L^2(D)} \; + \; k^2 \|\tilde{g}\|_{L^2(D)} \right] \quad \text{for all } \phi \in X \,. \end{aligned}$$

Ignoring for a moment the second integral on the left hand side of (5.63) let us find $v \in X$ with

$$(v, \phi)_X \;=\; \ell(\phi) \quad \text{for all } \phi \in X \,.$$

Were X a Hilbert space i.e., a complete inner product space, the Theorem A.31 of Riesz-Fischer would yield the existence of a unique element $r \in X$ with $(r, \phi)_X = \ell(\phi)$ for all $\phi \in X$. Since X, equipped with this norm given by (5.65) is **not** complete, we construct its completion with respect to that norm (see Theorem A.7 of the Appendix). In fact, the resulting complete inner product space is exactly the Sobolev space $H_0^1(D)$ defined after Definition A.20.

The functional ℓ can then be extended to a bounded functional on $H_0^1(D)$, and the Theorem A.31 of Riesz-Fischer yields the existence of a unique $r \in H_0^1(D)$ with $(r, \phi)_X = \ell(\phi)$ for all $\phi \in H_0^1(D)$. By the same arguments it can be shown that there exists a linear bounded operator K from $H_0^1(D)$ into itself with

$$\iint\limits_D v\,\overline{\phi}\,dx \;=\; (Kv, \phi)_X \quad \text{for all } v, \phi \in H_0^1(D) \,.$$

Therefore, the variational equation (5.63) is equivalent to

$$v \;-\; (k^2+1)Kv \;=\; r \quad \text{in } H_0^1(D) \,. \tag{5.66}$$

Using Rellich's imbedding theorem one can show that K is compact in $H_0^1(D)$. Therefore, the Theorem of Riesz (Theorem A.40 of the Appendix) is applicable to this equation. In particular, existence of a unique solution $v \in H_0^1(D)$ is guaranteed if the homogeneous equation (i.e. the equation for $r = 0$ or, equivalently, $g = 0$) admits only the trivial solution $v = 0$.

For the numerical treatment of the variational equation (5.63) one chooses a family of **ultimately dense subspaces** $X_n \subset H_0^1(D)$ (see Subsection 3.2.5) and determines $v_n \in X_n$ such that

$$\iint\limits_D \left(\nabla v_n \cdot \nabla\overline{\phi} - k^2 v_n\,\overline{\phi}\right) dx \;=\; -\iint\limits_D \left(\nabla\tilde{g} \cdot \nabla\overline{\phi} - k^2 \tilde{g}\,\overline{\phi}\right) dx \tag{5.67}$$

for all functions $\phi \in X_n$. By repeating the above arguments for X_n instead of $H_0^1(D)$ it can be shown that this variational equation can be written in the form

$$v_n \;-\; (k^2+1)\,P_n K v_n \;=\; P_n r \quad \text{in } X_n \,,$$

where $P_n : H_0^1(D) \longrightarrow X_n$ denotes the orthogonal projection operator onto X_n. This equation is uniquely solvable in X_n provided the homogeneous equation (5.63) admits only the trivial solution $v = 0$ [54]. This gives a general description of the variational approach.

In the finite element method itself, one usually chooses X_n to be a space of functions which enjoy a certain global smoothness property i.e., a subspace of some $C^p(\overline{D})$, and are locally polynomials of certain fixed order. To be more specific we start with a subdivision of the region D into non-overlapping quadrilaterals and/or triangles τ_i (with straight or curvilinear boundaries) which are called **finite elements**. We assume that

$$D \;=\; \bigcup_{i=1}^{n} \tau_i\,,$$

i.e., the elements form an exact partition of D. For given $p, q \in \mathbb{N}$ we define the **finite element spaces** $S_n^{p,q}$ by

$$S_n^{p,q} \;:=\; \bigl\{ v \in C^p(\overline{D}) : v|_{\tau_i} \in \mathcal{P}_q(\tau_i) \text{ for all } i = 1, \dots, n \bigr\}\,,$$

where $\mathcal{P}_q(\tau)$ denotes the space of polynomials in two independent variables of degree at most q on the set $\tau \subset \mathbb{R}^2$. By h we denote the maximal diameter of all elements τ_i. In the variational equation (5.67) we take $X_n = S_n^{p,q}$. The triangulation is assumed to satisfy the standard finite element geometrical constraints [24] and to be non-degenerate in the sense of Scott [119].

Example 5.27. As an example we consider a triangulation of D into n triangles with degree zero of smoothness and order one i.e.

$$S_n^{0,1} \;:=\; \bigl\{ v \in C(\overline{D}) : v|_{\tau_i} \text{ linear for every } i = 1, \dots, n \bigr\}\,.$$

We denote by $\boldsymbol{x}_1, \dots, \boldsymbol{x}_N$ those nodes of the triangulation which lie strictly in the interior D. Then $\dim X_n = N$, and a basis of X_n is given by the "hat functions" ϕ_j which vanish at all of the grid points except at $\boldsymbol{x}_j$ where $\phi_j(\boldsymbol{x}_j) = 1$. Making an ansatz of the form

$$v_n \;=\; \sum_{j=1}^{N} a_j\, \phi_j$$

for the solution $v_n \in S_n^{0,1}$ of the variation equation (5.67), and testing this equation with $\phi = \phi_\ell$, yields the finite algebraic system

$$\sum_{j=1}^{N} a_j \iint\limits_D \bigl(\nabla\phi_j \cdot \nabla\overline{\phi_\ell} - k^2 \phi_j\, \overline{\phi_\ell}\bigr)\, dx \;=\; -\iint\limits_D \bigl(\nabla\tilde{g} \cdot \nabla\overline{\phi_\ell} - k^2 \tilde{g}\, \overline{\phi_\ell}\bigr)\, dx \quad (5.68)$$

for all $\ell = 1, \dots, N$. We note that the coefficients

$$a_{\ell j} \;:=\; \iint\limits_{D} \bigl(\nabla\phi_j \cdot \nabla\overline{\phi_\ell} - k^2 \phi_j\,\overline{\phi_\ell}\bigr)\,dx\,, \quad \ell, j = 1, \ldots, N\,,$$

vanish for those ℓ and j whenever the supports of ϕ_j and ϕ_ℓ are disjoint. The hermitian matrix $(a_{\ell,j})$ is therefore a band-structured matrix which makes the numerical treatment of the linear system accessible by a number of fast algorithms. We note, however, that the matrix fails to be positive definite.

Now we turn to the *exterior* Dirichlet boundary value problem (5.2) in contrast to the treatment of the interior problem we have just discussed. The boundary data g is again given. The methods to be described below introduce an artificial domain containing Ω in its interior. For simplicity, we take this to be a disc of sufficiently large radius R and boundary $\Gamma_R = \{\boldsymbol{x} \in \mathbb{R}^2 : |\boldsymbol{x}| = R\}$. Again, we denote by $D_R := \{\boldsymbol{x} \in \Omega^c : |\boldsymbol{x}| < R\}$ the region between $\partial\Omega$ and Γ_R. The simplest approximation of the solution u of the exterior Dirichlet problem (5.2) is to replace the radiation condition (5.1c) by the boundary condition of impedance type

$$\frac{\partial u}{\partial r} \;-\; iku \;=\; 0 \quad \text{on } \Gamma_R\,. \tag{5.69}$$

Doing so leads to a boundary value problem in the bounded domain D_R. Its variational form is easily derived by multiplying the Helmholtz equation by $\overline{\phi(\boldsymbol{x})}$ for some $\phi \in C^1(\overline{D_R})$ with $\phi|_{\partial\Omega} = 0$, integrating over D_R, and using Green's first theorem. This results in

$$\iint\limits_{D_R} \bigl(\nabla u \cdot \nabla\overline{\phi} \;-\; k^2\, u\,\overline{\phi}\bigr)\,dx \;=\; -\int\limits_{|\boldsymbol{x}|=R} \overline{\phi}\,\frac{\partial u}{\partial r}\,d\ell\,. \tag{5.70}$$

Using the boundary condition (5.69) leads to

$$\iint\limits_{D_R} \bigl(\nabla u \cdot \nabla\overline{\phi} \;-\; k^2\, u\,\overline{\phi}\bigr)\,dx \;=\; -ik \int\limits_{|\boldsymbol{x}|=R} u\,\overline{\phi}\,d\ell\,. \tag{5.71}$$

It can be shown that the sesquilinear form

$$(u, \phi) \;\mapsto\; \int\limits_{|\boldsymbol{x}|=R} u\,\overline{\phi}\,d\ell$$

is bounded on $H^1(D_R) \times H^1(D_R)$ so that the variational form (5.71) is well-defined in the space $H^1_{0i}(D_R)$ which is the completion of the space $\bigl\{\phi \in C^1(\overline{D_R}) : \phi|_{\partial\Omega} = 0\bigr\}$ with respect to the norm $\|\phi\|_{H^1(D_R)}$. As above we extend g to a function $\tilde{g} \in H^1(D_R)$ with $\tilde{g} = 0$ on Γ_R and split u in the form $u = v + \tilde{g}$ with some $v \in H^1_{0i}(D_R)$. Then, (5.71) is transformed into

$$\iint\limits_{D_R} \left(\nabla v\cdot\nabla\overline{\phi} - k^2\, v\,\overline{\phi}\right) dx + \int\limits_{|x|=R} v\,\overline{\phi}\, d\ell \;=\; \iint\limits_{D_R} \left(\nabla\tilde{g}\cdot\nabla\overline{\phi} - k^2\, \tilde{g}\,\overline{\phi}\right) dx \qquad (5.72)$$

for all $\phi \in H^1_{0i}(D_R)$.

For the numerical approximation we proceed as in the approach above and replace $H^1_{0i}(D_R)$ by finite element spaces. Again, the order of convergence $u_n \to u$ depend on the smoothness of the solution u and the approximation properties of the finite element spaces, see e.g., [54].

Because of the slow decay of u and $\partial u/\partial r - iku$, considerable amount work has been published which seeks to improve the boundary condition (5.69) by using higher order approximations of the field. We do not want to go into details here and instead refer to the literature (see [54] for an overview).

A fairly recent and very successful approach is the **Perfectly Matched Layer Method** (PML method) in which the differential equation is changed outside of some disk $\{x : |x| \geq R\}$ in such a way that the solution $\tilde{u}$ of the new problem coincides with the original solution u inside this disc but decays exponentially as r tends to infinity. Then the simple Dirichlet boundary condition $\tilde{u} = 0$ on Γ_{R_1} for some $R_1 > R$ will lead to a good approximation of u. In our elementary presentation we follow the work of Collino and Monk ([28]). The basic idea of the PML method can be derived from the series expansion of the field u in the form (see Example 5.20)

$$u(r,\theta) \;=\; \sum_{n\in\mathbb{Z}} u_n\, H_n^{(1)}(kr)\, \mathrm{e}^{in\theta}\,, \quad r \geq R\,,\ 0 \leq \theta \leq 2\pi\,,$$

where R is again such that $\overline{\Omega}$ is contained in the open disc of radius R. We note that $H_n^{(1)}(kr)$ behaves asymptotically as

$$H_n^{(1)}(kr) \;=\; \sqrt{\frac{2}{\pi k}}\, \mathrm{e}^{-i\pi/4-in\pi/2}\, \frac{\exp(ikr)}{\sqrt{r}} \left[1 \,+\, \mathcal{O}\left(\frac{1}{r}\right)\right] \quad \text{as } r \,\to\, \infty\,,$$

(see (2.68a)). For $\operatorname{Im} k > 0$, this form shows exponential order of convergence as r tends to infinity (in contrast to the case where k is real). We choose $R_1 > R$ and an arbitrary real valued function $\psi \in C^1(\mathbb{R})$ with $\psi(s) > 0$ for $s \in (R, R_1)$ and $\psi(s) = 0$ for $s \notin [R, R_1]$. With the (complex) change of variable

$$\rho \;=\; \rho(r) \;=\; r \,+\, i\,r \int\limits_R^r \psi(s)\, ds\,, \quad r > 0\,,$$

we observe that $\rho(r) = r$ for $r \leq R$ and $\rho(r) = \alpha r$ for $r \geq R_1$ where $\alpha := 1 + i\int_R^{R_1} \psi(s)\, ds$ has positive imaginary part. Defining

$$\tilde{u}(r,\theta) \;=\; \begin{cases} \displaystyle\sum_{n\in\mathbb{Z}} u_n\, H_n^{(1)}\big(k\rho(r)\big)\, \mathrm{e}^{in\theta}\,, & r \geq R\,,\ 0 \leq \theta \leq 2\pi\,, \\ u(r,\theta)\,, & r < R\,,\ 0 \leq \theta \leq 2\pi\,, \end{cases} \qquad (5.73)$$

we observe that $\tilde{u}$ coincides with u for $r \le R$ and

$$\tilde{u}(r,\theta) \;=\; \sum_{n\in\mathbb{Z}} u_n\, H_n^{(1)}(k\alpha r)\, e^{in\theta} \quad \text{for } r \ge R_1\,,\ 0 \le \theta \le 2\pi\,. \tag{5.74}$$

In contrast to u, the function $\tilde{u}$ decays exponentially to zero as r tends to infinity since $\operatorname{Im}\alpha > 0$. It remains to derive the differential equation for $\tilde{u}$. In polar coordinates (ρ, θ) the original Helmholtz equation takes the form

$$\frac{1}{\rho}\frac{\partial}{\partial\rho}\left(\rho\frac{\partial u}{\partial\rho}\right) + \frac{1}{\rho^2}\frac{\partial^2 u}{\partial\theta^2} + k^2 u \;=\; 0\,.$$

With the substitution $\rho = \rho(r)$ we have

$$\frac{\partial}{\partial\rho} = \frac{1}{\rho'(r)}\frac{\partial}{\partial r}\,.$$

Therefore, the Helmholtz equation in terms of $\tilde{u}$ takes the form

$$\frac{1}{\rho\,\rho'}\frac{\partial}{\partial r}\left(\frac{\rho}{\rho'}\frac{\partial\tilde{u}}{\partial r}\right) + \frac{1}{\rho^2}\frac{\partial^2\tilde{u}}{\partial\theta^2} + k^2\tilde{u} \;=\; 0\,,$$

i.e.

$$\frac{1}{r}\frac{\partial}{\partial r}\left(a(r)\,r\frac{\partial\tilde{u}}{\partial r}\right) + \frac{1}{a(r)\,r^2}\frac{\partial^2\tilde{u}}{\partial\theta^2} + \frac{k^2\rho(r)\,\rho'(r)}{r}\tilde{u} \;=\; 0\,, \tag{5.75}$$

where $a(r) = \rho(r)/\bigl(r\,\rho'(r)\bigr)$. We can write this also in Cartesian coordinates using

$$\frac{\partial}{\partial r} \;=\; \cos\theta\frac{\partial}{\partial x_1} + \sin\theta\frac{\partial}{\partial x_2} \quad\text{and}\quad \frac{\partial}{\partial\theta} \;=\; -r\sin\theta\frac{\partial}{\partial x_1} + r\cos\theta\frac{\partial}{\partial x_2}$$

as

$$\operatorname{div}\bigl(A(\boldsymbol{x})\,\nabla\tilde{u}(\boldsymbol{x})\bigr) + k^2\,c(\boldsymbol{x})\,\tilde{u}(\boldsymbol{x}) \;=\; 0\,, \tag{5.76}$$

where

$$A(\boldsymbol{x}) = \left.\begin{pmatrix} a(r)\cos^2\theta + \sin^2\theta/a(r) & \cos\theta\sin\theta\bigl[a(r) - 1/a(r)\bigr] \\ \cos\theta\sin\theta\bigl[a(r) - 1/a(r)\bigr] & a(r)\sin^2\theta + \cos^2\theta/a(r) \end{pmatrix}\right|_{(r,\theta)=\boldsymbol{x}} \quad\text{and}$$

$$c(\boldsymbol{x}) = \frac{\rho(|\boldsymbol{x}|)\,\rho'(|\boldsymbol{x}|)}{|\boldsymbol{x}|}\,.$$

So far, the exterior boundary value problem (5.2) and the problem of finding $\tilde{u}$ which satisfies (5.76) in Ω^c with $\tilde{u} = g$ on $\partial\Omega$, and which has an expansion in the form of (5.74) are completely equivalent. Now we find an *approximate* solution by choosing $R_2 > R_1$, defining the region $D_{R_2} := \bigl\{\boldsymbol{x} \in \Omega^c : |\boldsymbol{x}| < R_2\bigr\}$, and considering the Dirichlet problem in D_{R_2} with boundary data $\tilde{u} = g$ on $\partial\Omega$ and $\tilde{u} = 0$ on Γ_{R_2}. This problem can easily be solved by any standard finite element package. For more details, in particular error estimates as $R_2 \to \infty$, we refer to [28, 80].

5.5.4 Hybrid Methods

The so-called hybrid methods also reduce the Dirichlet boundary value problem for the unbounded domain Ω^c to one on the bounded domain D_R but take a different approach to incorporating the influence of the field exterior to Γ_R. We start again with the variational equation (5.70). To incorporate the radiation condition, we introduce the **Dirichlet-to-Neumann operator** Λ as follows:

For every $\psi \in C^\infty(\Gamma_R)$ let u_ψ be the unique solution of the exterior Dirichlet problem in $\mathbb{R}^2 \setminus \overline{D_R}$ with boundary condition $u_\psi = \psi$ on Γ_R. By Example 5.20 u_ψ has the expansion

$$u_\psi(r,\theta) \;=\; \sum_{n\in\mathbb{Z}} \frac{a_n}{H_n^{(1)}(kR)}\, H_n^{(1)}(kr)\, e^{in\theta}\,, \quad r > R,\; 0 \le \theta \le 2\pi\,,$$

where $\psi(\theta) = \sum_{n\in\mathbb{Z}} a_n \exp(in\theta)$. We define $\Lambda\psi$ as the normal derivative of u_ψ on Γ_R which we can compute explicitly:

$$(\Lambda\psi)(\theta) \;:=\; \left.\frac{\partial u_\psi(r,\theta)}{\partial r}\right|_{r=R} \;=\; k \sum_{n\in\mathbb{Z}} a_n \,\frac{(H_n^{(1)})'(kR)}{H_n^{(1)}(kR)}\, e^{in\theta}\,, \quad 0 \le \theta \le 2\pi\,.$$

Certainly, if u is a radiating solution of the Helmholtz equation in $\mathbb{R}^2 \setminus \overline{\Omega}$ then $\partial u/\partial r\big|_{\Gamma_R} = \Lambda u|_{\Gamma_R}$. We substitute this result into (5.70) and find that

$$\iint\limits_{D_R} \left(\nabla u \cdot \nabla\overline{\phi} \;-\; k^2\, u\,\overline{\phi}\right) dx \;=\; -\int\limits_{|x|=R} \overline{\phi}\,\Lambda u \, d\ell \tag{5.77}$$

for all ϕ with $\phi|_{\partial\Omega} = 0$. It can be shown (see [69]) that the sesquilinear form

$$(\phi, u) \;\mapsto\; \int\limits_{|x|=R} \overline{\phi}\,\Lambda u\, d\ell$$

is bounded on $H^1(D_R) \times H^1(D_R)$ so that the variational form (5.77) is well defined in the space $H^1_{0i}(D_R)$ which is the completion of the space $\{\phi \in C^1(\overline{D_R}) \,:\, \phi|_{\partial\Omega} = 0\}$ with respect to the norm $\|\phi\|_{H^1(D_R)}$. As above we extend g to a function $\tilde{g} \in H^1(D_R)$ with $\tilde{g} = 0$ on Γ_R and split u in the form $u = v + \tilde{g}$ with some $v \in H^1_{0i}(D_R)$. Then, (5.77) is transformed into

$$\iint\limits_{D_R} \left(\nabla v \cdot \nabla\overline{\phi} \;-\; k^2\, v\,\overline{\phi}\right) dx \;+\; \int\limits_{|x|=R} \overline{\phi}\,\Lambda v\, d\ell \;=\; \iint\limits_{D_R} \left(\nabla \tilde{g} \cdot \nabla\overline{\phi} \;-\; k^2\, \tilde{g}\,\overline{\phi}\right) dx$$

for all $\phi \in H^1_{0i}(D_R)$.

For the numerical approximation we proceed as in the approach above and replace $H^1_{0i}(D_R)$ by finite element spaces $S_h \subset H^1_{0i}(D_R)$. In addition, we truncate the Dirichlet-to-Neumann operator and define

$$(\Lambda_N\psi)(\theta) \;:=\; \frac{k}{2\pi}\sum_{|n|\le N}\big(g,\exp(in\cdot)\big)_{L^2(\Gamma_R)}\,\frac{(H_n^{(1)})'(kR)}{H_n^{(1)}(kR)}\,\mathrm{e}^{in\theta}\,,\quad 0\le\theta\le 2\pi\,.$$

The finite dimensional problem then is to determine $v_{h,N}\in S_h$ such that

$$\iint\limits_{D_R}\big(\nabla v_{h,N}\cdot\nabla\overline{\phi}-k^2\,v_{h,N}\,\overline{\phi}\big)\,dx+\int\limits_{|x|=R}\overline{\phi}\,\Lambda_N v_{h,N}\,d\ell \;=\;\iint\limits_{D_R}\big(\nabla\tilde{g}\cdot\nabla\overline{\phi}-k^2\,\tilde{g}\,\overline{\phi}\big)\,dx$$

for all $\phi\in S_h$. For a detailed error analysis we refer to [69].

The main problem in implementing this particular method is to compute the non-local boundary condition in (5.77) with sufficient accuracy. In addition, the unusual boundary conditions means that standard finite element packages cannot be used easily, and the matrix $A=\big(a_{\ell j}\big)$ for the discrete problem contains a dense submatrix. The following method avoids these disadvantages by decoupling the interior and the exterior problem using an auxiliary function ψ on Γ_R which has then to be determined by demanding approximate continuity of u and $\partial u/\partial r$ on Γ_R. The method, is motivated by the work of Bramble and Pasciak [19, 20] on Laplace's equation. It shares the advantages of the Bramble and Pasciak method and allows us both to use a simple Hankel function expansion of the scattered field outside D_R and to use any suitable standard finite element method to approximate the problem in D_R. For example, the iterative techniques of [17] can be used to implement a fast Helmholtz equation solver on the bounded domain D_R. Another advantage of the analysis is that, following the ideas of Bramble [19], we can suggest an efficient method for matching the solutions across Γ_R using the conjugate gradient method.

The analysis carried out in [69] for the case of an inhomogeneous medium shows that there is no stability constraint involving the mesh size h for the interior finite elements and the order N of the exterior Hankel function basis. This is unexpected in view of the analysis of Bramble [19] and of Bramble and Pasciak [20] but is due to the special spaces we use.

To illustrate this method we consider the following boundary value problems for D_R and $\mathbb{R}^2\setminus\overline{D_R}$:

Given $\psi\in C(\partial\Omega)$, determine the solution $w_i=w_i(\psi)$ of

$$\Delta w_i\;+\;k^2 w_i=0\quad\text{in } D_R\,,\tag{5.78a}$$

$$w_i=0\quad\text{on }\partial\Omega\,,\tag{5.78b}$$

$$\frac{\partial w_i}{\partial r}\;+\;ik\,w_i=\psi\quad\text{on }\Gamma_R,\tag{5.78c}$$

and the solution $w_e=w_e(\psi)$ of

$$\Delta w_e\;+\;k^2 w_e=0\quad\text{in }\mathbb{R}^2\setminus\overline{D_R}\,,\tag{5.79a}$$

$$\frac{\partial w_e}{\partial r}\;+\;ik\,w_e=\psi\quad\text{on }\Gamma_R\,,\tag{5.79b}$$

$$\frac{\partial w_e}{\partial r}\;-\;ik\,w_e=\mathcal{O}\big(r^{-3/2}\big)\quad\text{as } r\to\infty\,.\tag{5.79c}$$

Furthermore, we define the function v by the solution of the problem

$$\Delta v \; + \; k^2 v = 0 \quad \text{in } D_R\,, \tag{5.80a}$$

$$v = f \text{ on } \partial\Omega, \tag{5.80b}$$

$$\frac{\partial v}{\partial r} \; + \; ik\, v = 0 \quad \text{on } \Gamma_R\,. \tag{5.80c}$$

The solutions have to be understood in the variational sense, as in the case of the Dirichlet problem above. For example, multiplication of (5.78a) by some $\overline{\phi} \in C^1(\overline{D_R})$ with $\phi|_{\partial\Omega} = 0$, integration over D_R, and applying Green's first formula yields

$$\iint\limits_D \left(\nabla w_i \cdot \nabla \overline{\phi} - k^2 w_i \,\overline{\phi}\right) dx \; + \; ik \int\limits_{\Gamma_R} w_i \,\overline{\phi}\, d\ell \;=\; \int\limits_{\Gamma_R} \psi \,\overline{\phi}\, d\ell\,. \tag{5.81}$$

which defines the variational solution w_i if we allow the solution w_i and the test function ϕ to be in the Sobolev space $H^1_{0i}(D_R)$ defined above. The solution v of (5.80a)–(5.80c) is defined analogously by

$$\iint\limits_D \left(\nabla v \cdot \nabla \overline{\phi} - k^2 v \,\overline{\phi}\right) dx \; + \; ik \int\limits_{\Gamma_R} v \,\overline{\phi}\, d\ell \;=\; 0 \tag{5.82}$$

for all $\phi \in H^1_{0i}(D_R)$.

Rather than using the variational formulation for the exterior problem (5.79a)–(5.79c), we will instead use the fact that the concepts of the variational solution and parallel curves as presented in Section 5.4 are equivalent. Therefore, by Example 5.20, the solution w_e has the representation in the form

$$w_e(r,\theta) \;=\; \frac{1}{k} \sum_{n\in\mathbb{Z}} \frac{a_n\, e^{in\theta}}{(H_n^{(1)})'(kR) + iH_n^{(1)}(kR)}\, H_n^{(1)}(kr)\,, \quad r > R\,,\; 0 \le \theta \le 2\pi\,, \tag{5.83}$$

where $\psi(\theta) = \sum\limits_{n=1}^{\infty} a_n \exp(in\theta)$.

We then define the function $u = u(\psi)$ by

$$u(\boldsymbol{x}) \;=\; \begin{cases} w_e(\boldsymbol{x})\,, & \text{for } |\boldsymbol{x}| > R\,, \\ w_i(\boldsymbol{x}) + v(\boldsymbol{x})\,, & \text{for } \boldsymbol{x} \in D_R\,. \end{cases} \tag{5.84}$$

We remark that, from our construction, $\partial u/\partial r + iku$ is continuous across Γ_R. Thus we need to choose ψ to make u continuous across Γ_R, and so it is easy to see that u solves the exterior Dirichlet boundary value problem (5.2) provided ψ satisfies the equation:

$$w_e(\psi) \;-\; w_i(\psi) \;=\; v \quad \text{on } \Gamma_R\,. \tag{5.85}$$

Analogous arguments as in [69] prove the equivalence of this equation with the boundary value problem (5.2). In particular, this approach can be used to show existence and uniqueness of H^1–solutions of (5.2).

For the approximate computation of w_i, we use a standard finite element space on a triangulation of D_R as introduced above. By h we denote again the maximal diameter of all elements τ_i. Let $S_h \subset H^1(D_R)$ denote the finite element space of those elements ϕ_h with $\phi_h|_{\partial\Omega} = 0$.

Now we define $w_i^h \in S_h$ and $v^h \in S_h$ as the usual finite element solution of (5.81), (5.82), e.g.

$$\iint\limits_D \left(\nabla w_i^h \cdot \nabla\overline{\phi_h} - k^2 w_i^h \overline{\phi_h}\right) dx \; + \; ik \int\limits_{\Gamma_R} w_i^h \overline{\phi_h}\, d\ell \; = \; \int\limits_{\Gamma_R} \psi \overline{\phi_h}\, d\ell \tag{5.86}$$

for all $\phi_h \in S_h$. The element v^h is defined analogously.

For the approximate computation of w_e we truncate the series representation (5.83). This is equivalent to taking ψ from the subspace

$$\dot{S}_N \; := \; \left\{ \psi \in L^2(\Gamma_R) : \psi(\theta) = \sum_{n=-N}^{N} a_n \, \mathrm{e}^{in\theta}, \; a_n \in \mathbb{C} \right\}. \tag{5.87}$$

for some $N \in \mathbb{N}$. For $\psi \in \dot{S}_N$ we compute w_e explicitly as

$$w_e(r,\theta) \; = \; \frac{1}{k} \sum_{n=-N}^{N} \frac{a_n \, \mathrm{e}^{in\theta}}{(H_n^{(1)})'(kR) + iH_n^{(1)}(kR)} \, H_n^{(1)}(kr) \, ,$$

for $r > R$ and $0 \le \theta \le 2\pi$. In a manner analogous to the definition of $u(\psi)$ we define the function u^h by

$$u^h(\boldsymbol{x}) \; = \; \begin{cases} w_e(\boldsymbol{x}) \, , & \text{for } |\boldsymbol{x}| > R \, , \\ w_i^h(\boldsymbol{x}) + v^h(\boldsymbol{x}) \, , & \text{for } \boldsymbol{x} \in D_R \, . \end{cases}$$

and take ψ from the finite dimensional space $\dot{S}_N$. Equation (5.85) is then replaced by the finite dimensional problem to determine $\psi \in \dot{S}_N$ such that

$$w_e \; - \; P_N w_i^h \; = \; P_N v^h \quad \text{on } \Gamma_R \tag{5.88}$$

where $P_N : L^2(\Gamma_R) \to \dot{S}_N$ is the orthogonal projection onto $\dot{S}_N$. For more details on this method, in particular error estimates in scales of Sobolev spaces, we refer to the original paper [70].

6

Boundary Value Problems for Maxwell's Equations

This chapter is devoted to the mathematical investigation of a particular boundary value problem for Maxwell's equation. We consider the time harmonic case, i.e. the equations (2.13a)–(2.13d), for the homogeneous and non-conducting case, i.e. when the permittivity and permeability, ε and μ, are constant and the conductivity $\sigma = 0$. We restrict ourselves to the boundary condition $\boldsymbol{n} \times \boldsymbol{H} = \boldsymbol{h}$ on $\partial\Omega$. For more complicated situations, in particular for the Leontovich (or impedance) and conductive boundary conditions we refer to [29], [5].

The organization of this chapter is similar to the preceeding one. We begin with a suitable representation theorem for solutions of Maxwell's equation, show uniqueness of radiating solutions and present an outline of the classical existence theory through the systematic use of vector potentials. The power radiated at infinity is given by the far field pattern as we have seen in 2.35. The unique solvability of the exterior boundary value problem *defines* the far field operator F which maps the boundary data onto the far field pattern. At the end of this chapter we show that this operator can be extended to on operator acting on L^2-functions defined on the boundary $\partial\Omega$. This makes it possible to apply the results on existence and uniqueness of optimization problems of Chapter 3.

6.1 Introduction and Formulation of the Problem

Let $\boldsymbol{E}$ and $\boldsymbol{H}$ denote the electric and magnetic field, respectively, for the time harmonic case in vacuo. As we showed in (2.13a)–(2.13d), these fields satisfy Maxwell's equation in the form

$$\begin{aligned} \operatorname{curl} \boldsymbol{H} &= -i\omega\varepsilon\, \boldsymbol{E} \\ \operatorname{curl} \boldsymbol{E} &= i\omega\mu \boldsymbol{H} \end{aligned}$$

The equations $\operatorname{div} \boldsymbol{E} = 0$ and $\operatorname{div} \boldsymbol{H} = 0$ follow immediately from these equations since we consider only homogeneous media.

These field equations must be satisfied outside of the domain Ω which denotes the radiating structure. On boundary $\partial\Omega$ of Ω we impose the boundary condition

$$\boldsymbol{n} \times \boldsymbol{H} \;=\; \boldsymbol{h} \quad \text{on} \quad \partial\Omega\,,$$

where $\boldsymbol{n}(\boldsymbol{x})$ denotes the unit normal vector at $\boldsymbol{x} \in \partial\Omega$, directed into the exterior of Ω and $\boldsymbol{h}$ is the electric current density.[1]

In addition, $\boldsymbol{E}$ and $\boldsymbol{H}$ have to satisfy the Silver-Müller radiation condition

$$\boldsymbol{E} \times \hat{\boldsymbol{x}} + \frac{1}{Y_0}\boldsymbol{H} \;=\; \mathcal{O}\left(\frac{1}{r^2}\right), \quad r \to \infty\,,$$

with admittance $Y_0 = \sqrt{\varepsilon/\mu}$. Solutions of the field equations which satisfy this asymptotic condition will be called *radiating solutions* or *radiating fields.*

For the precise mathematical setting we need spaces of vector fields on $\partial\Omega$. By $C_T(\partial\Omega)$ we denote the space of continuous tangential vector fields i.e.,

$$C_T(\partial\Omega) \;=\; \left\{\boldsymbol{a} : \partial\Omega \to \mathbb{C}^3 : a_j \in C(\partial\Omega),\ j = 1,2,3,\ \boldsymbol{a}\cdot\boldsymbol{n} = 0 \ \text{ on } \ \partial\Omega\right\}.$$

Given a tangential field $\boldsymbol{h} \in C_T(\partial\Omega)$ the boundary value problem is to find the vector fields $\boldsymbol{E}, \boldsymbol{H} \in C^1(\mathbb{R}^3 \setminus \overline{\Omega}) \cap C(\mathbb{R}^3 \setminus \Omega)$ which satisfy

$$\operatorname{curl}\boldsymbol{H} \;+\; i\omega\varepsilon\,\boldsymbol{E} = \boldsymbol{o} \quad \text{in} \quad \mathbb{R}^3 \setminus \overline{\Omega}\,, \tag{6.1a}$$

$$\operatorname{curl}\boldsymbol{E} \;-\; i\omega\mu\boldsymbol{H} = \boldsymbol{o} \quad \text{in} \quad \mathbb{R}^3 \setminus \overline{\Omega}\,, \tag{6.1b}$$

$$\boldsymbol{n} \times \boldsymbol{H} \;=\; \boldsymbol{h} \quad \text{on} \quad \partial\Omega\,, \tag{6.1c}$$

$$\boldsymbol{E} \times \hat{\boldsymbol{x}} + \frac{1}{Y_0}\boldsymbol{H} \;=\; \mathcal{O}\left(\frac{1}{r^2}\right), \quad r \to \infty\,, \quad \text{uniformly with respect to } \hat{\boldsymbol{x}} = \boldsymbol{x}/r. \tag{6.1d}$$

Before we study this boundary value problem in more detail, we recall the Stratton-Chu formula (see [29, 53]). Let $k := \omega\sqrt{\mu\varepsilon}$ be the wave number for the case $\sigma = 0$, and

$$\Phi(\boldsymbol{x},\boldsymbol{y}) \;:=\; \frac{\exp(ik|\boldsymbol{x}-\boldsymbol{y}|)}{4\pi|\boldsymbol{x}-\boldsymbol{y}|}\,, \quad \boldsymbol{x} \neq \boldsymbol{y},$$

be the fundamental solution of the 3-dimensional Helmholtz equation. Then we have

Theorem 6.1. *Assume that $\Omega \in \mathbb{R}^3$ is a bounded domain with C^2-boundary $\partial\Omega$ and with unit normal vector $\boldsymbol{n}$ on $\partial\Omega$ directed into the exterior of Ω. Let $\boldsymbol{E}, \boldsymbol{H} \in C^1(\mathbb{R}^3 \setminus \overline{\Omega}) \cap C(\mathbb{R}^3 \setminus \Omega)$ be a radiating solution of the Maxwell equations*

[1] For convenience we will assume, throughout, that the boundary is C^2 although much weaker assumptions can be made (see e.g. [100] who treats Lipschitz surfaces).

$$\begin{aligned} \operatorname{curl}\boldsymbol{H} \;+\; i\omega\varepsilon\,\boldsymbol{E} &= \boldsymbol{o} \quad \text{in} \quad \mathbb{R}^3\setminus\overline{\Omega}\,,\\ \operatorname{curl}\boldsymbol{E} \;-\; i\omega\mu\boldsymbol{H} &= \boldsymbol{o} \quad \text{in} \quad \mathbb{R}^3\setminus\overline{\Omega}\,. \end{aligned}$$

Then

$$\begin{aligned} \boldsymbol{E}(\boldsymbol{x}) = \operatorname{curl}\int_{\partial\Omega} \boldsymbol{n}(\boldsymbol{y})\times\boldsymbol{E}(\boldsymbol{y})\,\Phi(\boldsymbol{x},\boldsymbol{y})\,dS(\boldsymbol{y}) \\ -\frac{1}{i\omega\varepsilon}\operatorname{curl}^2\int_{\partial\Omega} \boldsymbol{n}(\boldsymbol{y})\times\boldsymbol{H}(\boldsymbol{y})\,\Phi(\boldsymbol{x},\boldsymbol{y})\,dS(\boldsymbol{y})\,, \end{aligned} \tag{6.2a}$$

$$\begin{aligned} \boldsymbol{H}(\boldsymbol{x}) = \operatorname{curl}\int_{\partial\Omega} \boldsymbol{n}(\boldsymbol{y})\times\boldsymbol{H}(\boldsymbol{y})\,\Phi(\boldsymbol{x},\boldsymbol{y})\,dS(\boldsymbol{y}) \\ +\frac{1}{i\omega\mu}\operatorname{curl}^2\int_{\partial\Omega} \boldsymbol{n}(\boldsymbol{y})\times\boldsymbol{E}(\boldsymbol{y})\,\Phi(\boldsymbol{x},\boldsymbol{y})\,dS(\boldsymbol{y}) \end{aligned} \tag{6.2b}$$

for $\boldsymbol{x}\in\mathbb{R}^3\setminus\overline{\Omega}$.

Proof: Define the modified fields $\tilde{\boldsymbol{E}}$ and $\tilde{\boldsymbol{H}}$ by $\tilde{\boldsymbol{E}} = \sqrt{\varepsilon}\,\boldsymbol{E}$ and $\tilde{\boldsymbol{H}} = \sqrt{\mu}\,\boldsymbol{H}$. Then $\tilde{\boldsymbol{E}}$ and $\tilde{\boldsymbol{H}}$ satisfy $\operatorname{curl}\tilde{\boldsymbol{E}} - ik\tilde{\boldsymbol{H}} = 0$ and $\operatorname{curl}\tilde{\boldsymbol{H}} + ik\tilde{\boldsymbol{E}} = 0$ and the radiation condition

$$\tilde{\boldsymbol{E}}\times\hat{\boldsymbol{x}} \;+\; \tilde{\boldsymbol{H}} \;=\; \mathcal{O}\left(\frac{1}{r^2}\right), \quad r\to\infty.$$

In this symmetric form, the Stratton-Chu formula has been proven in, e.g., Colton/Kress[29]:

$$\begin{aligned} \tilde{\boldsymbol{E}}(\boldsymbol{x}) &= \operatorname{curl}\int_{\partial\Omega} \boldsymbol{n}(\boldsymbol{y})\times\tilde{\boldsymbol{E}}(\boldsymbol{y})\,\Phi(\boldsymbol{x},\boldsymbol{y})\,dS(\boldsymbol{y}) \\ &\quad -\frac{1}{ik}\operatorname{curl}^2\int_{\partial\Omega} \boldsymbol{n}(\boldsymbol{y})\times\tilde{\boldsymbol{H}}(\boldsymbol{y})\,\Phi(\boldsymbol{x},\boldsymbol{y})\,dS(\boldsymbol{y})\,,\\ \tilde{\boldsymbol{H}}(\boldsymbol{x}) &= \operatorname{curl}\int_{\partial\Omega} \boldsymbol{n}(\boldsymbol{y})\times\tilde{\boldsymbol{H}}(\boldsymbol{y})\,\Phi(\boldsymbol{x},\boldsymbol{y})\,dS(\boldsymbol{y}) \\ &\quad +\frac{1}{ik}\operatorname{curl}^2\int_{\partial\Omega} \boldsymbol{n}(\boldsymbol{y})\times\tilde{\boldsymbol{E}}(\boldsymbol{y})\,\Phi(\boldsymbol{x},\boldsymbol{y})\,dS(\boldsymbol{y})\,, \end{aligned}$$

$x\notin\overline{\Omega}$. Substituting the form of $\tilde{\boldsymbol{E}}$ and $\tilde{\boldsymbol{H}}$ yields the assertion. □

From this formula we draw the following conclusions:

Every component E_j and H_j of the fields satisfies the scalar Helmholtz equation

$$\Delta u \;+\; k^2 u \;=\; 0 \quad \text{in} \quad \mathbb{R}^3 \setminus \overline{\Omega}\,.$$

In particular, the fields $\boldsymbol{E}$ and $\boldsymbol{H}$ are analytic functions in $\mathbb{R}^3 \setminus \overline{\Omega}$. Also, we can derive the far field patterns of $\boldsymbol{E}$ and $\boldsymbol{H}$ from the asymptotic form of the fundamental solution Φ (compare with (2.42a)):

$$\Phi(\boldsymbol{x},\boldsymbol{y}) \;=\; \frac{\mathrm{e}^{ikr}}{4\pi\, r}\,\mathrm{e}^{-ik\hat{\boldsymbol{x}}\cdot\boldsymbol{y}} \;+\; \mathcal{O}\left(\frac{1}{r^2}\right), \quad r = |\boldsymbol{x}| \to \infty\,,$$

uniformly with respect to $\boldsymbol{y} \in \partial\Omega$ and $\hat{\boldsymbol{x}} = \boldsymbol{x}/r \in S^2$. We have:

Theorem 6.2. *Let the assumption of the previous theorem be satisfied. Then* $\boldsymbol{E}$ *and* $\boldsymbol{H}$ *have the form*

$$\boldsymbol{E}(\boldsymbol{x}) = \frac{\mathrm{e}^{ikr}}{r}\,\boldsymbol{E}_\infty(\hat{\boldsymbol{x}}) \;+\; \mathcal{O}\left(\frac{1}{r^2}\right), \quad r \to \infty\,, \tag{6.3a}$$

$$\boldsymbol{H}(\boldsymbol{x}) = Y_0\,\frac{\mathrm{e}^{ikr}}{r}\,\boldsymbol{H}_\infty(\hat{\boldsymbol{x}}) \;+\; \mathcal{O}\left(\frac{1}{r^2}\right), \quad r \to \infty\,, \tag{6.3b}$$

uniformly with respect to $\hat{\boldsymbol{x}} = \boldsymbol{x}/r \in S^2$*, where*

$$\boldsymbol{E}_\infty(\hat{\boldsymbol{x}}) = \frac{ik}{4\pi}\,\hat{\boldsymbol{x}} \times \int\limits_{\partial\Omega} \left\{ [\boldsymbol{n}(\boldsymbol{y}) \times \boldsymbol{E}(\boldsymbol{y})] + \frac{1}{Y_0}\,[(\boldsymbol{n}(\boldsymbol{y}) \times \boldsymbol{H}(\boldsymbol{y})) \times \hat{\boldsymbol{x}}] \right\} \mathrm{e}^{-ik\hat{\boldsymbol{x}}\cdot\boldsymbol{y}}\, dS(\boldsymbol{y})\,,$$

$$\boldsymbol{H}_\infty(\hat{\boldsymbol{x}}) = \frac{ik}{4\pi}\,\hat{\boldsymbol{x}} \times \int\limits_{\partial\Omega} \left\{ \frac{1}{Y_0}\,[\boldsymbol{n}(by) \times \boldsymbol{H}(\boldsymbol{y}) - [(\boldsymbol{n}(\boldsymbol{y}) \times \boldsymbol{E}(\boldsymbol{y}) \times \hat{\boldsymbol{x}}] \right\} \mathrm{e}^{-ik\hat{\boldsymbol{x}}\cdot\boldsymbol{y}}\, dS(\boldsymbol{y})\,,$$

for $\hat{\boldsymbol{x}} \in S^2$*. Furthermore,* $\boldsymbol{H}_\infty = \hat{\boldsymbol{x}} \times \boldsymbol{E}_\infty$ *and* $\hat{\boldsymbol{x}} \cdot \boldsymbol{E}_\infty = \hat{\boldsymbol{x}} \cdot \boldsymbol{H}_\infty = 0$.

Proof: For a constant vector $\boldsymbol{a} \in \mathbb{C}^3$ we derive

$$\operatorname{curl}_x\big[\boldsymbol{a}\,\Phi(\boldsymbol{x},\boldsymbol{y})\big] \;=\; \frac{ik}{4\pi}\,\frac{\mathrm{e}^{ikr}}{r}\left\{\mathrm{e}^{-ik\hat{\boldsymbol{x}}\cdot\boldsymbol{y}}(\hat{\boldsymbol{x}} \times \boldsymbol{a}) \;+\; \mathcal{O}\left(\frac{1}{r}\right)\right\} \tag{6.4a}$$

$$\operatorname{curl}_x^2\big[\boldsymbol{a}\,\Phi(\boldsymbol{x},\boldsymbol{y})\big] \;=\; \frac{k^2}{4\pi}\,\frac{\mathrm{e}^{ikr}}{r}\left\{\mathrm{e}^{-ik\hat{\boldsymbol{x}}\cdot\boldsymbol{y}}[\hat{\boldsymbol{x}} \times (\boldsymbol{a} \times \hat{\boldsymbol{x}})] \;+\; \mathcal{O}\left(\frac{1}{r}\right)\right\} \tag{6.4b}$$

as $r = |\boldsymbol{x}| \to \infty$ uniformly with respect to $\boldsymbol{y} \in \partial\Omega$ and $\hat{\boldsymbol{x}} = \boldsymbol{x}/r \in S^2$. Substituting (6.4a), (6.4b) into (6.2a), (6.2b) the conclusion follows easily. □

6.2 Uniqueness and Existence

The following theorem is the basis of the uniqueness theorem. For a proof we refer to [30].

Theorem 6.3. *Assume that $\Omega \in \mathbb{R}^3$ is a bounded domain with C^2-boundary $\partial\Omega$ such the unit normal vector $\boldsymbol{n}$ on $\partial\Omega$ is directed into the exterior. Let $\boldsymbol{E}, \boldsymbol{H} \in C^1(\mathbb{R}^3 \setminus \overline{\Omega}) \cap C(\mathbb{R}^3 \setminus \Omega)$ be radiating solutions of the Maxwell equations*

$$\begin{aligned} \operatorname{curl} \boldsymbol{H} \;+\; i\omega\varepsilon\, \boldsymbol{E} &= \boldsymbol{o} \quad \text{in} \quad \mathbb{R}^3 \setminus \overline{\Omega}, \\ \operatorname{curl} \boldsymbol{E} \;-\; i\omega\mu \boldsymbol{H} &= \boldsymbol{o} \quad \text{in} \quad \mathbb{R}^3 \setminus \overline{\Omega}. \end{aligned}$$

If

$$\operatorname{Re} \int_{\partial\Omega} [\boldsymbol{n}(\boldsymbol{y}) \times \boldsymbol{E}(\boldsymbol{y})] \cdot \overline{\boldsymbol{H}(\boldsymbol{y})}\, dS(\boldsymbol{y}) \;\leq\; 0 \tag{6.5}$$

then $\boldsymbol{E} \equiv \boldsymbol{H} \equiv 0$ in $\mathbb{R}^3 \setminus \Omega$.

From this we conclude immediately

Theorem 6.4. *There exists at most one solution of the boundary value problem (6.1).*

Now we turn to the problem of existence. Motivated by the representation (6.2a), (6.2b) we make the assumption that $\boldsymbol{H}$ has the form

$$\boldsymbol{H}(\boldsymbol{x}) = \operatorname{curl} \int_{\partial\Omega} \boldsymbol{a}(\boldsymbol{y})\, \Phi(\boldsymbol{x}, \boldsymbol{y})\, dS(\boldsymbol{y}) + i\eta \operatorname{curl}^2 \int_{\partial\Omega} \boldsymbol{n}(\boldsymbol{y}) \times (S_0^2 \boldsymbol{a})(\boldsymbol{y})\, \Phi(\boldsymbol{x}, \boldsymbol{y})\, dS(\boldsymbol{y}) \tag{6.6}$$

for some density $\boldsymbol{a} \in C_T(\partial\Omega)$ and a real parameter $\eta \in \mathbb{R}$. Here, S_0 denotes the single layer operator

$$S_0\varphi(\boldsymbol{x}) \;=\; \int_{\partial\Omega} \frac{1}{4\pi|\boldsymbol{x} - \boldsymbol{y}|}\, \varphi(\boldsymbol{y})\, dS(\boldsymbol{y}), \quad \boldsymbol{x} \in \partial\Omega, \tag{6.7}$$

corresponding to $k = 0$. We note that S_0 is selfadjoint in $L^2(\partial\Omega)$, and maps $C^{0,\alpha}(\partial\Omega)$ compactly into $C^{1,\alpha}(\partial\Omega)$ (see Theorem 5.14 which holds also in $\mathbb{R}^3$). In the equation (6.6) we apply the operator S_0^2 to each component a_j of $\boldsymbol{a}$.

The choice of the correct space of functions $\boldsymbol{a} \in C_T(\partial\Omega)$ is more delicate than in the scalar case. In particular it is necessary to introduce the concept of the **surface divergence** of tangential vector fields. For a continuously differentiable (scalar) function φ on $\partial\Omega$ the surface gradient $\operatorname{Grad}\varphi \in C_T(\partial\Omega)$ is defined by the (uniquely existing) tangential vector field $\operatorname{Grad}\varphi \in C_T(\partial\Omega)$ with

$$\lim_{t \to 0} \frac{1}{t} \left|\varphi(\boldsymbol{x} + t\boldsymbol{h}(\boldsymbol{x})) - \varphi(\boldsymbol{x}) - t \operatorname{Grad}\varphi(\boldsymbol{x}) \cdot \boldsymbol{h}(\boldsymbol{x})\right| \;=\; 0$$

for every vector field $\boldsymbol{h} \in C_T(\partial\Omega)$ (cf. Section 5.3). We define the surface divergence by Gauss' theorem:

Definition 6.5. *A tangential vector field $\boldsymbol{a} \in C_T(\partial\Omega)$ has a* **weak surface divergence**, *if there exists a continuous scalar function,* $\operatorname{Div} \boldsymbol{a} \in C(\partial\Omega)$, *such that*

$$\int_{\partial\Omega} \varphi \operatorname{Div} \boldsymbol{a} \, dS \;=\; -\int_{\partial\Omega} \operatorname{Grad} \varphi \cdot \boldsymbol{a} \, dS \tag{6.8}$$

for all $\varphi \in C^1(\partial\Omega)$.

It is easily seen by a denseness argument that the surface divergence is unique if it exists.

For any continuous vector field $\boldsymbol{E}$ defined in a neighborhood $U = \{\boldsymbol{x} + t\boldsymbol{n}(\boldsymbol{x}) : \boldsymbol{x} \in \partial\Omega,\ 0 < t < t_0\}$ for which the normal component of $\operatorname{curl} \boldsymbol{E}$ is continuous in $\overline{U}$ it can be shown (see [29]) that the surface divergence of $\boldsymbol{n} \times \boldsymbol{E}$ exists and is given by

$$\operatorname{Div}(\boldsymbol{n} \times \boldsymbol{E}) \;=\; -\boldsymbol{n} \cdot \operatorname{curl} \boldsymbol{E} \quad \text{on } \partial\Omega\,. \tag{6.9}$$

We define the space $C_D(\partial\Omega)$ and the corresponding space of Hölder continuous functions (see Appendix, Section A.2) by

$$C_D(\partial\Omega) = \{\boldsymbol{a} \in C_T(\partial\Omega) : \operatorname{Div} \boldsymbol{a} \in C(\partial\Omega)\}\,, \tag{6.10a}$$

$$C_D^{0,\alpha}(\partial\Omega) = \{\boldsymbol{a} \in C_T^{0,\alpha}(\partial\Omega) : \operatorname{Div} \boldsymbol{a} \in C^{0,\alpha}(\partial\Omega)\} \tag{6.10b}$$

with their canonical norms

$$\|\boldsymbol{a}\|_{C_D(\partial\Omega)} = \|\boldsymbol{a}\|_{C_T(\partial\Omega)} \;+\; \|\operatorname{Div} \boldsymbol{a}\|_{C(\partial\Omega)}\;,$$
$$\|\boldsymbol{a}\|_{C_D^{0,\alpha}(\partial\Omega)} = \|\boldsymbol{a}\|_{C_T^{0,\alpha}(\partial\Omega)} \;+\; \|\operatorname{Div} \boldsymbol{a}\|_{C^{0,\alpha}(\partial\Omega)}\;.$$

Before we use the equation (6.6) to prove existence, we state the jump conditions for the vector potential (see [29]):

Theorem 6.6. *For given $\boldsymbol{a} \in C_T^{0,\alpha}(\partial\Omega)$ we define the vector potential*

$$\boldsymbol{F}(\boldsymbol{x}) \;:=\; \int_{\partial\Omega} \boldsymbol{a}(\boldsymbol{y})\, \Phi(\boldsymbol{x},\boldsymbol{y})\, dS(\boldsymbol{y})\,, \quad \boldsymbol{x} \notin \partial\Omega\,. \tag{6.11}$$

Then $\boldsymbol{F}$ and its first derivatives can be uniformly Hölder-continuously extended from Ω to $\overline{\Omega}$ and from $\mathbb{R}^3 \setminus \overline{\Omega}$ to $\mathbb{R}^3 \setminus \Omega$ with limiting values

$$\boldsymbol{F}(\boldsymbol{x})|_\pm = \int_{\partial\Omega} \boldsymbol{a}(\boldsymbol{y})\, \Phi(\boldsymbol{x},\boldsymbol{y})\, dS(\boldsymbol{y})\,, \quad \boldsymbol{x} \in \partial\Omega\,, \tag{6.12a}$$

$$\operatorname{div} \boldsymbol{F}(\boldsymbol{x})|_\pm = \int_{\partial\Omega} \nabla_x \Phi(\boldsymbol{x},\boldsymbol{y}) \cdot \boldsymbol{a}(\boldsymbol{y})\, dS(\boldsymbol{y})\,, \quad \boldsymbol{x} \in \partial\Omega\,, \tag{6.12b}$$

$$\operatorname{curl} \boldsymbol{F}(\boldsymbol{x})|_\pm = \mp\frac{1}{2}\, \boldsymbol{n}(\boldsymbol{x}) \times \boldsymbol{a}(\boldsymbol{x}) + \int_{\partial\Omega} \nabla_x \Phi(\boldsymbol{x},\boldsymbol{y}) \times \boldsymbol{a}(\boldsymbol{y})\, dS(\boldsymbol{y})\,, \tag{6.12c}$$

$\boldsymbol{x} \in \partial\Omega$, where $\boldsymbol{F}(\boldsymbol{x})|_\pm = \lim\limits_{\varepsilon\to 0} \boldsymbol{F}\big(\boldsymbol{x} \pm \varepsilon \boldsymbol{n}(\boldsymbol{x})\big)$. Analogous relations hold for $\operatorname{div} \boldsymbol{F}(\boldsymbol{x})|_\pm$ and $\operatorname{curl} \boldsymbol{F}(\boldsymbol{x})|_\pm$.

For $\boldsymbol{a} \in C^{0,\alpha}_D(\partial\Omega)$, $\operatorname{curl}^2 \boldsymbol{F}$ can also be uniformly Hölder-continuously extended from Ω to $\overline{\Omega}$ and from $\mathbb{R}^3 \setminus \overline{\Omega}$ to $\mathbb{R}^3 \setminus \Omega$ with limiting values

$$\operatorname{curl}^2 \boldsymbol{F}(\boldsymbol{x})|_\pm \;=\; \mp \frac{1}{2}\, \boldsymbol{n}(\boldsymbol{x}) \operatorname{Div} \boldsymbol{a}(\boldsymbol{x}) + k^2\, \boldsymbol{F}(\boldsymbol{x}) + \int\limits_{\partial\Omega} \nabla_x \Phi(\boldsymbol{x},\boldsymbol{y}) \operatorname{Div} \boldsymbol{a}(\boldsymbol{y})\, dS(\boldsymbol{y}), \tag{6.12d}$$

$\boldsymbol{x} \in \partial\Omega$. Furthermore, there exists $c > 0$ with

$$\begin{aligned} \|\boldsymbol{F}\|_{C^{0,\alpha}(\overline{G})} &\le c\, \|\boldsymbol{a}\|_{C^{0,\alpha}(\partial\Omega)} , \\ \|\operatorname{div} \boldsymbol{F}\|_{C^{0,\alpha}(\overline{G})} &\le c\, \|\boldsymbol{a}\|_{C^{0,\alpha}(\partial\Omega)} , \\ \|\operatorname{curl} \boldsymbol{F}\|_{C^{0,\alpha}(\overline{G})} &\le c\, \|\boldsymbol{a}\|_{C^{0,\alpha}(\partial\Omega)} , \\ \left\|\operatorname{curl}^2 \boldsymbol{F}\right\|_{C^{0,\alpha}(\overline{G})} &\le c\, \|\boldsymbol{a}\|_{C^{0,\alpha}_D(\partial\Omega)} , \end{aligned}$$

where $G = \Omega$ or $G = \mathbb{R}^3 \setminus \overline{\Omega}$.

A proof of (6.12a)–(6.12c) can be found in [30]. For $\operatorname{curl}^2 \boldsymbol{F}$ we note that, for $\boldsymbol{x} \notin \partial\Omega$,

$$\operatorname{curl}^2 \boldsymbol{F} \;=\; -\Delta \boldsymbol{F} + \nabla \operatorname{div} \boldsymbol{F} \;=\; k^2 \boldsymbol{F} + \nabla \operatorname{div} \boldsymbol{F}$$

and

$$\begin{aligned} \operatorname{div} \boldsymbol{F}(\boldsymbol{x}) &= \int\limits_{\partial\Omega} \nabla_x \Phi(\boldsymbol{x},\boldsymbol{y}) \cdot \boldsymbol{a}(\boldsymbol{y})\, dS(\boldsymbol{y}) \;=\; -\int\limits_{\partial\Omega} \nabla_y \Phi(\boldsymbol{x},\boldsymbol{y}) \cdot \boldsymbol{a}(\boldsymbol{y})\, dS(\boldsymbol{y}) \\ &= \int\limits_{\partial\Omega} \Phi(\boldsymbol{x},\boldsymbol{y}) \operatorname{Div} \boldsymbol{a}(\boldsymbol{y})\, dS(\boldsymbol{y}) . \end{aligned}$$

The jump conditions (5.17b) for the derivatives of the single layer potential yield equation (6.12d). □

We note that for $\boldsymbol{a} \in C^{0,\alpha}_D(\partial\Omega)$

$$\boldsymbol{n}(\boldsymbol{x}) \times \operatorname{curl} \boldsymbol{F}(\boldsymbol{x})|_\pm \;=\; \pm \frac{1}{2}\, \boldsymbol{a}(\boldsymbol{x}) + \boldsymbol{n}(\boldsymbol{x}) \times \int\limits_{\partial\Omega} \nabla_x \Phi(\boldsymbol{x},\boldsymbol{y}) \times \boldsymbol{a}(\boldsymbol{y})\, dS(\boldsymbol{y}) \tag{6.13a}$$

and

$$\boldsymbol{n}(\boldsymbol{x}) \times \operatorname{curl}^2 \boldsymbol{F}(\boldsymbol{x})|_\pm \;=\; k^2\, \boldsymbol{n}(\boldsymbol{x}) \times \boldsymbol{F}(\boldsymbol{x}) + \boldsymbol{n}(\boldsymbol{x}) \times \int\limits_{\partial\Omega} \nabla_x \Phi(\boldsymbol{x},\boldsymbol{y}) \operatorname{Div} \boldsymbol{a}(\boldsymbol{y})\, dS(\boldsymbol{y}), \tag{6.13b}$$

for $x \in \partial\Omega$. The right hand sides of these two relations define two boundary operators $M, N : C^{0,\alpha}_D(\partial\Omega) \longrightarrow C^{0,\alpha}_D(\partial\Omega)$ by

$$(M\boldsymbol{a})(\boldsymbol{x}) := \boldsymbol{n}(\boldsymbol{x}) \times \int_{\partial\Omega} \nabla_x \Phi(\boldsymbol{x},\boldsymbol{y}) \times \boldsymbol{a}(\boldsymbol{y})\, dS(\boldsymbol{y}), \tag{6.14a}$$

$$(N\boldsymbol{a})(\boldsymbol{x}) := k^2\, \boldsymbol{n}(\boldsymbol{x}) \times \int_{\partial\Omega} \Phi(\boldsymbol{x},\boldsymbol{y}) \boldsymbol{a}(\boldsymbol{y})\, dS(\boldsymbol{y}) \tag{6.14b}$$
$$+ \boldsymbol{n}(\boldsymbol{x}) \times \int_{\partial\Omega} \nabla_x \Phi(\boldsymbol{x},\boldsymbol{y})\, \operatorname{Div} \boldsymbol{a}(\boldsymbol{y})\, dS(\boldsymbol{y})$$

for $x \in \partial\Omega$. Both are well-defined and bounded. The operator M is even compact as is proven in [30]. (Note that the operator N used here corresponds to the map $\boldsymbol{a} \mapsto N(\boldsymbol{a} \times \boldsymbol{n})$ in [30].)

The representation of $\boldsymbol{H}$ in equation (6.6) involves two vector potentials with densities $\boldsymbol{a} \in C_D^{0,\alpha}(\partial\Omega)$ and $i\eta\, \boldsymbol{n} \times S_0^2 \boldsymbol{a}$, respectively. Therefore, (6.6) satisfies the boundary condition (6.1c) if and only if $\boldsymbol{a} \in C_D^{0,\alpha}(\partial\Omega)$ satisfies the equation

$$\frac{1}{2}\boldsymbol{a} \;+\; M\boldsymbol{a} \;+\; i\eta N\big(\boldsymbol{n} \times S_0^2 \boldsymbol{a}\big) \;=\; \boldsymbol{h}. \tag{6.15}$$

From the mapping properties of S_0^2 (see Theorem 5.14) we note that $\boldsymbol{a} \mapsto \boldsymbol{n} \times S_0^2 \boldsymbol{a}$ is compact from $C_D^{0,\alpha}(\partial\Omega)$ into itself. Therefore, the equation (6.15) is of the form

$$\boldsymbol{a} \;+\; K\boldsymbol{a} \;=\; 2\boldsymbol{h}$$

with some compact operator $K : C_D^{0,\alpha}(\partial\Omega) \longrightarrow C_D^{0,\alpha}(\partial\Omega)$. By the **Riesz theory**, existence follows from uniqueness, see Theorem A.40. To prove uniqueness[2] assume that $\boldsymbol{a} \in C_D^{0,\alpha}(\partial\Omega)$ solves (6.15) for $\boldsymbol{h} = \boldsymbol{o}$. Define $\boldsymbol{H}$ by (6.6) and $\boldsymbol{E}$ by $\boldsymbol{E} = -\frac{1}{\omega\varepsilon} \operatorname{curl} \boldsymbol{H}$. Then $\boldsymbol{n} \times \boldsymbol{H} = \boldsymbol{o}$ on $\partial\Omega$. The uniqueness result yields $\boldsymbol{H} \equiv \boldsymbol{E} \equiv \boldsymbol{o}$ in $\mathbb{R}^3 \setminus \Omega$. From Theorem 6.6 we conclude that

$$\begin{aligned} \boldsymbol{n} \times \boldsymbol{H}|_- &= \boldsymbol{n} \times \boldsymbol{H}|_- \;-\; \boldsymbol{n} \times \boldsymbol{H}|_+ \;=\; -\boldsymbol{a}, \\ \boldsymbol{n} \times \operatorname{curl} \boldsymbol{H}|_- &= \boldsymbol{n} \times \operatorname{curl} \boldsymbol{H}|_- \;-\; \boldsymbol{n} \times \operatorname{curl} \boldsymbol{H}|_+ \;=\; -i\eta\, k^2\, \boldsymbol{n} \times S_0^2 \boldsymbol{a}. \end{aligned}$$

Hence, from Gauss' theorem we have

$$\begin{aligned} i\eta\, k^2 \int_{\partial\Omega} |S_0 \boldsymbol{a}|^2 dS = i\eta\, k^2 \int_{\partial\Omega} \overline{\boldsymbol{a}} \cdot S_0^2 \boldsymbol{a}\, dS \;&=\; \int_{\partial\Omega} \big[\boldsymbol{n} \times \overline{\boldsymbol{H}}|_-\big] \cdot \operatorname{curl} \boldsymbol{H}\, dS \\ &= \iint_{\Omega} \big[|\operatorname{curl} \boldsymbol{H}|^2 - k^2 |\boldsymbol{H}|^2\big]\, dx. \end{aligned}$$

Taking the imaginary part yields $S_0 \boldsymbol{a} = \boldsymbol{o}$ and hence $\boldsymbol{a} = \boldsymbol{o}$ since S_0 is one-to-one. Thus we have shown that equation (6.15) has a unique solution

[2] We emphasize that here, as earlier when we treated the scalar case, we are concerned with the uniqueness of solutions of the *integral equation* (6.15) and *not* with the unique solvability of the original boundary value problem (6.1).

$\boldsymbol{a} \in C_D^{0,\alpha}(\partial\Omega)$ for every $\boldsymbol{h} \in C_D^{0,\alpha}(\partial\Omega)$. In other words, the operator $\frac{1}{2}I + M + i\eta N(\boldsymbol{n} \times S_0^2)$ is an isomorphism from $C_D^{0,\alpha}(\partial\Omega)$ onto itself.

From the equation (6.6) and the asymptotic forms (6.4a), (6.4b) we observe that the far field pattern $\boldsymbol{H}_\infty$ of $\boldsymbol{H}$ is given by

$$\begin{aligned}\boldsymbol{H}_\infty(\hat{\boldsymbol{x}}) &= \frac{1}{4\pi Y_0}\hat{\boldsymbol{x}} \times \int\limits_{\partial\Omega} \boldsymbol{a}(\boldsymbol{y})\, \mathrm{e}^{-ik\,\hat{\boldsymbol{x}}\cdot\boldsymbol{y}}\, dS(\boldsymbol{y}) \\ &+ \frac{i\eta}{4\pi Y_0}\hat{\boldsymbol{x}} \times \int\limits_{\partial\Omega} \left[\boldsymbol{n}(\boldsymbol{y}) \times (S_0^2\boldsymbol{a})(\boldsymbol{y})\right] \times \hat{\boldsymbol{x}}\, \mathrm{e}^{-ik\,\hat{\boldsymbol{x}}\cdot\boldsymbol{y}}\, dS(\boldsymbol{y}), \quad \hat{\boldsymbol{x}} \in S^2 .\end{aligned} \tag{6.16}$$

The operator $\boldsymbol{a} \mapsto \boldsymbol{H}_\infty$ from $C_D^{0,\alpha}(\partial\Omega)$ into $C_T(S^2)$ is certainly compact. The preceeding analysis can be summarized in the following theorem:

Theorem 6.7. *For every $\boldsymbol{h} \in C_D^{0,\alpha}(\partial\Omega)$ the boundary value problem (6.1a)–(6.1d) has a unique solution $\boldsymbol{E}, \boldsymbol{H} \in C^1(\mathbb{R}^3 \setminus \overline{\Omega}) \cap C(\mathbb{R}^3 \setminus \Omega)$. The operator $\mathcal{K} : C_D^{0,\alpha}(\partial\Omega) \longrightarrow C_T(S^2)$, $\boldsymbol{h} \mapsto \boldsymbol{H}_\infty$, is well-defined and compact. Here, $\boldsymbol{H}_\infty$ is the far field pattern corresponding to the solution of the boundary value problem (6.1a)–(6.1d) with boundary values $\boldsymbol{h}$.*

6.3 L^2–Boundary Data

As we have seen in Chapter 3, for optimization problems it is desirable to work in Hilbert spaces of functions. It is our aim to extend the compact operator $\mathcal{K} : C_D^{0,\alpha}(\partial\Omega) \to C_T(S^2)$ to one from $L_T^2(\partial\Omega)$ into $C_T(S^2)$. Here, $L_T^2(\partial\Omega)$ is defined by

$$L_T^2(\partial\Omega) \;=\; \left\{\boldsymbol{a} : \partial\Omega \to \mathbb{C}^3 : a_j \in L^2(\partial\Omega),\ j = 1,2,3,\ \boldsymbol{a}\cdot\boldsymbol{n} = 0 \text{ a.e. on } \partial\Omega\right\}.$$

As we have seen in the previous section the operator $\mathcal{K}$ is the composition of the operators $\left[\frac{1}{2}I + M + i\eta N(\boldsymbol{n} \times S_0^2)\right]^{-1}$ and $\boldsymbol{a} \mapsto \boldsymbol{H}_\infty$ where $\boldsymbol{H}_\infty$ is given by the form (6.16). The integral in (6.16) is also well defined for L^2–vector fields $\boldsymbol{a} \in L_T^2(\partial\Omega)$, and $\boldsymbol{a} \mapsto \boldsymbol{H}_\infty$ is certainly compact from $L_T^2(\partial\Omega)$ into $C_T(S^2)$. Therefore, it suffices to show that $\left[\frac{1}{2}I + M + i\eta N(\boldsymbol{n} \times S_0^2)\right]^{-1}$ has a bounded extension in $L_T^2(\partial\Omega)$. To show this we will again use Theorem 5.18 due to Lax. We apply that theorem to show that M and $N(\boldsymbol{n} \times S_0^2)$ are compact in $L_T^2(\partial\Omega)$. Let us first consider M. By changing the orders of integration we see that the adjoint operator M' of M with respect to the bilinear form

$$\langle \boldsymbol{a}, \boldsymbol{b}\rangle \;:=\; \int\limits_{\partial\Omega} \boldsymbol{a}\cdot\boldsymbol{b}\, dS$$

is given by $M' = RMR$ where $R\boldsymbol{a} = \boldsymbol{a}\times\boldsymbol{n}$. Since M is compact in $C_T^{0,\alpha}(\partial\Omega)$ so is M' and thus also the L^2–adjoint $M^* = \overline{M}'$ where $\overline{M}\boldsymbol{a} = \overline{M\overline{\boldsymbol{a}}}$. Application of Theorem 5.18 to $X = Y = C_T^{0,\alpha}(\partial\Omega)$ yields that M is also compact in $L^2_T(\partial\Omega)$.

The operator N consists of two parts. The first part is $-k^2 RS$ which is compact in $C_T(\partial\Omega)$ by Theorem 5.11 and thus in $L^2_T(\partial\Omega)$ by Theorem 5.18. To study the second part we consider the auxiliary operator

$$(K\varphi)(\boldsymbol{x}) \;:=\; \boldsymbol{n}(\boldsymbol{x}) \times \int_{\partial\Omega} \nabla_x \Phi(\boldsymbol{x},\boldsymbol{y})\, \varphi(\boldsymbol{y})\, dS(\boldsymbol{y}), \quad \boldsymbol{x} \in \partial\Omega .$$

By Theorem 5.10 this operator is bounded from $C^{0,\alpha}(\partial\Omega)$ into $C_T^{0,\alpha}(\partial\Omega)$. Its adjoint with respect to $\langle\cdot,\cdot\rangle$ is given by

$$(K'\varphi)(\boldsymbol{x}) \;:=\; \int_{\partial\Omega} \big[\boldsymbol{n}(\boldsymbol{y}) \times \boldsymbol{a}(\boldsymbol{y})\big] \cdot \nabla_y \Phi(\boldsymbol{x},\boldsymbol{y})\, dS(\boldsymbol{y}), \quad \boldsymbol{x} \in \partial\Omega ,$$

which is, by Theorem 6.6, bounded from $C_T^{0,\alpha}(\partial\Omega)$ into $C^{0,\alpha}(\partial\Omega)$. Therefore, K^* is also bounded and application of Theorem 5.18 to $X = C^{0,\alpha}(\partial\Omega)$ and $Y = C_T^{0,\alpha}(\partial\Omega)$ yields that K is bounded from $L^2(\partial\Omega)$ to $L^2_T(\partial\Omega)$.

Since the single layer operator, S_0, is compact in $L^2_T(\partial\Omega)$ it remains to prove that the operator $A : \boldsymbol{a} \mapsto \operatorname{Div}\big(\boldsymbol{n} \times S_0\boldsymbol{a}\big)$ is bounded from $L^2_T(\partial\Omega)$ into $L^2(\partial\Omega)$. From the mapping properties of S_0 we observe that A is bounded from $C_T^{0,\alpha}(\partial\Omega)$ into $C^{0,\alpha}(\partial\Omega)$. Then using $\operatorname{Div}(\boldsymbol{n} \times \boldsymbol{u}) = -\boldsymbol{n} \cdot \operatorname{curl} \boldsymbol{u}$ and (6.12c) we conclude that

$$(A\boldsymbol{a})(\boldsymbol{x}) \;=\; \operatorname{Div}\big(\boldsymbol{n}\times S_0\boldsymbol{a}\big)(\boldsymbol{x}) \;=\; -\boldsymbol{n}(\boldsymbol{x})\cdot \int_{\partial\Omega} \nabla_x \Phi(\boldsymbol{x},\boldsymbol{y}) \times \boldsymbol{a}(\boldsymbol{y})\, dS(\boldsymbol{y}), \quad \boldsymbol{x} \in \partial\Omega .$$

Changing the orders of integration yields

$$(A'\varphi)(\boldsymbol{x}) \;=\; \int_{\partial\Omega} \nabla_y \Phi(\boldsymbol{x},\boldsymbol{y}) \times \boldsymbol{n}(\boldsymbol{y})\, \varphi(\boldsymbol{y})\, dS(\boldsymbol{y}), \quad \boldsymbol{x} \in \partial\Omega .$$

This operator is bounded from $C^{0,\alpha}(\partial\Omega)$ into $C_T^{0,\alpha}(\partial\Omega)$ since $A' - K$ is compact. Finally, application of Theorem 5.18 yields that the operator A has a bounded extension from $L^2_T(\partial\Omega)$ into $L^2(\partial\Omega)$. Summarizing the results we have shown the first part of the following theorem:

Theorem 6.8. *The operator $\mathcal{K} : C_D^{0,\alpha}(\partial\Omega) \longrightarrow C_T(S^2)$ has a bounded extension from $L^2_T(\partial\Omega)$ into $C_T(\partial\Omega)$. It is the composition of the isomorphism $\left[\frac{1}{2}\, I + M + i\eta\, N(\boldsymbol{n} \times S_0^2)\right]^{-1}$ in $L^2_T(\partial\Omega)$ and the compact operator $\boldsymbol{a} \mapsto \boldsymbol{H}_\infty$ from $L^2_T(\partial\Omega)$ into $C_T(\partial\Omega)$. Furthermore, the range of $\mathcal{K}$ is dense in $C_T(S^2)$ and consists of analytic functions on S^2.*

Proof: It remains to prove the second part of the theorem concerning density of the range. But this is shown exactly the same way as in the proof of Theorem 5.19. One replaces the trigonometric sum $\sum_{n=-N}^{N} a_n e^{int}$ by the sum of spherical harmonics $\sum_{n=0}^{N} \sum_{m=-n}^{n} a_{nm} Y_n^m(\theta,\phi)$ and argues as before. □

7

Some Particular Optimization Problems

In this chapter, we will study several particular optimization problems which are of interest in antenna design. Of course, one such problem is the synthesis problem that we have discussed in Chapter 4. Here, we will treat (for the most part) problems in which the objective functionals describe intrinsic characteristics of the far field pattern as, for example, gain, directivity, or measures of efficiency, rather than the problem of finding the best approximation to a given far field pattern.

Thus we will discuss such problems as the problem of maximizing power in a given sector (or perhaps in a fixed direction) under various types of constraints, and treat this problem for some concrete cases of continuous sources. We also treat the problem of optimizing the signal-to-noise ratio of an antenna, as well as a particular special case what we term the "null-placement" problem, in which we attempt to constrain side-lobes in particular directions while optimizing some appropriate measure of performance for the antenna. As concrete examples of the application of the general results, we will present the particular cases of the finite line source and the circular loop.

7.1 General Assumptions

While we have studied quite general optimization problems in Chapter 3 and the application of the general techniques developed there to synthesis problems in Chapter 4, the present section is devoted to the maximization of the radiated power with respect to the surface current which we again denote by ψ. The general results of Chapter 3 are specialized here in that the optimality criteria $\mathcal{J}$ as well as the constraint set U is specified. It follows, in particular, that the general results of Chapter 3 guarantee the existence of an optimal solution for this problem. However, due to the particular nature of the performance criterion, we are able to derive some interesting *characterizations* of the optimal surface currents depending on the choice of the constraint sets U and it is this aspect of the theory upon which we concentrate.

Let the elements, ψ, from a complex Hilbert space X, describe the feeding of the antenna. As in the Chapter 3, our notation does not distinguish between scalar and vector quantities. We assume only that the operator $\mathcal{K} : X \to C(S^{d-1})$ has the following properties:

(A1) $\mathcal{K} : X \to C(S^{d-1})$ is compact and one-to-one. In particular, $\mathcal{K}$ is not identically zero.
(A2) $\mathcal{K}\psi$ is an analytic function on S^{d-1} for every $\psi \in X$.

Furthermore, we will take $U \subset X$ to be a non-empty, bounded, closed, and convex set and the function $\alpha \in L^\infty(S^{d-1})$ to be a real-valued, and non-negative with the property:

(A3) The support of α, i.e. the closed set

$$\mathcal{A} := \bigcap \{A \subset S^{d-1} : A \text{ closed and } \alpha = 0 \text{ a.e. on } S^{d-1} \setminus A\}$$

contains an open set (relative to S^{d-1}).

We may think of α being the characteristic function of some subset $\mathcal{A} \subset S^{d-1}$ with positive measure. These three assumptions imply that for $\psi \neq 0$ the analytic function $\mathcal{K}\psi$ cannot even vanish on the support of α. As in Section 3.4 we define the radiated power in a sector by[1]

$$\mathcal{J}_1(\psi) := \int\limits_{S^{d-1}} \alpha(\hat{x})^2 \, |(\mathcal{K}\psi)(\hat{x})|^2 \, ds(\hat{x}) = \|\alpha\mathcal{K}\psi\|^2_{L^2(S^{d-1})}, \quad \psi \in X, \tag{7.1}$$

and consider the optimization problem

$$\text{Maximize} \quad \mathcal{J}_1(\psi) = \|\alpha\mathcal{K}\psi\|^2_{L^2(S^{d-1})} \quad \text{subject to } \psi \in U. \tag{7.2}$$

We will also consider the maximization of

$$\mathcal{J}_2(\psi) := \sum_{j=1}^{m} w_j \, |(\mathcal{K}\psi)(\hat{x}_j)|^2, \quad \psi \in X, \tag{7.3}$$

subject to $\psi \in U$ where $w_j > 0$ are some positive weight factors.

Solutions $\psi^o \in U$ of these optimization problems exist according to the general existence Theorem 3.1 since U is weakly sequentially compact (Theorem 3.7) and $\mathcal{J}_1$, $\mathcal{J}_2$ are weakly sequentially continuous (Theorem 3.30). We note that the optimal value $\mathcal{J}_1(\psi^o)$ is positive whenever $U \neq \{0\}$. We make the general assumption that the optimal values are positive.

Our present interest is in the problem of finding characterizations of the optimal solutions which will be helpful in actual computations. An important property of $\mathcal{J}_1$ which is useful in reaching that goal is its strict convexity, i.e.

[1] In this section we write ds for both the differential dS in the case of surface integrals in $\mathbb{R}^3$ and for $d\ell$ in the case of line integrals in $\mathbb{R}^2$.

$$\mathcal{J}_1\big(\lambda\psi_1+(1-\lambda)\psi_2\big) \;<\; \lambda\mathcal{J}_1(\psi_1) + (1-\lambda)\mathcal{J}_1(\psi_2) \quad \text{for all } \psi_1 \neq \psi_2,\ \lambda \in (0,1)\,.$$

We summarize some of the most important properties of these functionals by recalling the result of Theorem 3.32:

Lemma 7.1. *Let $\mathcal{K} : X \to C(S^{d-1})$ and $\alpha \in L^\infty(S^{d-1})$ satisfy the assumptions (A1), (A2), and (A3), and let $f \in L^2(S^{d-1})$.*

(a) The functional

$$\mathcal{J}_1(\psi) \;:=\; \|\alpha\mathcal{K}\psi - f\|^2_{L^2(S^{d-1})}\,, \qquad \psi \in X\,,$$

is strictly convex and continuously Fréchet differentiable with gradient

$$\nabla\mathcal{J}_1(\psi) \;=\; 2\,\mathcal{K}^*(\alpha^2\mathcal{K}\psi - \alpha f)\,, \qquad \psi \in X\,,$$

where $\mathcal{K}^* : L^2(S^{d-1}) \to X$ denotes the adjoint of the operator $\mathcal{K}$ considered as an operator from X into $L^2(S^{d-1})$.

(b) The functional

$$\mathcal{J}(\psi) \;:=\; |(\mathcal{K}\psi)(\hat{\boldsymbol{x}})|^2\,, \qquad \psi \in X\,, \qquad \textit{with } \hat{\boldsymbol{x}} \in S^{d-1} \textit{ fixed,}$$

is convex and continuously Fréchet differentiable with gradient

$$\nabla\mathcal{J}(\psi) \;=\; 2\,(\mathcal{K}\psi)(\hat{\boldsymbol{x}})\,p$$

where $p \in X$ denotes the Riesz representation of the linear functional $\varphi \mapsto (\mathcal{K}\varphi)(\hat{\boldsymbol{x}})$, $\varphi \in X$, i.e. the unique element $p \in X$ with

$$(\mathcal{K}\varphi)(\hat{\boldsymbol{x}}) \;=\; \big(\varphi, p\big)_X \quad \textit{for all } \varphi \in X\,. \tag{7.4}$$

7.2 Maximization of Power

We start with a problem which is closely related to maximizing directivity, namely that of maximizing the radiated power in a preassigned angular sector of the far field. This sector is described by introducing the characteristic function, α, of a patch on the unit sphere. As we shall see, for simple power constraints this problem leads to an eigenvalue problem for its far field operator $\mathcal{K}$ and is accessible to numerical treatment by introducing appropriate approximation finite-dimensional problems.

As in our general discussion in Chapter 3, it is possible, at least formally, to consider the problem of maximizing the radiated power in one or more discrete directions, $\hat{\boldsymbol{x}}_1, \ldots, \hat{\boldsymbol{x}}_m$, as a special case by using δ-functions. The reader may wish to consult Section 3.4 as background. Again, we remark that a rigorous mathematical treatment would involve the use of distributions in the sense of L. Schwartz which is beyond the scope of the present book.

Our discussion begins with a consideration of the constrained problem with a simple power constraints on the input functions.

7.2.1 Input Power Constraints

We begin by taking the set U to be of the form

$$U \;:=\; \{\psi \in X : g(\psi) \;\leq\; 0\} \tag{7.5}$$

where $g : X \to \mathbb{R}$ is some continuous and uniformly convex function. Then we know from Lemma 3.9 that U is closed, convex and bounded.[2] Recall that we have identified the extreme points of U as just the boundary points of U in Lemma 3.18.

For the particular choice of U given by (7.5) the optimization problem (7.2) becomes

$$\text{Maximize} \quad \|\alpha\,\mathcal{K}\psi\|^2_{L^2(S^{d-1})} \ \text{ subject to } \psi \in X \text{ and } g(\psi) \leq 0\,.$$

Since the set of extreme points coincides with the boundary of the constraint set, the optimal solutions ψ^o necessarily satisfy $g(\psi^o) \;=\; 0$ by Theorem 3.16. Therefore, we can apply the Lagrange multiplier rule of Theorem 3.22 provided the function g is continuously Fréchet differentiable and $\nabla g(\psi^o) \neq 0$.

As an example we consider $g(\psi) = \|\psi\|^2_X - 1$ and observe, from the binomial theorem, that

$$g(\psi + \varphi) \;-\; g(\psi) \;=\; 2\,\mathrm{Re}\,(\psi, \varphi)_X \;+\; \|\varphi\|^2_X$$

and thus $\nabla g(\psi) = 2\psi$.

The application of the Lagrange multiplier rule, under the current hypothesis that $\alpha \in L^\infty(S^{d-1})$, then insures that for any optimal solution ψ^o of this optimization problem with $\nabla g(\psi^o) \neq 0$ there exists $\lambda \in \mathbb{R}$, $\lambda \geq 0$, (the *Lagrange multiplier*) such that

$$-\mathcal{K}^*(\alpha^2\mathcal{K}\psi^o) \;+\; \lambda\,\nabla g(\psi^o) \;=\; 0\,. \tag{7.6}$$

The particular example $U = \{\psi \in X : \|\psi\|_X \leq 1\}$ is frequently met in practice as it has the interpretation of limiting input power to the antenna. For this important case, (7.6) leads (after replacing λ by $\lambda/2$) to the *eigenvalue problem*

$$\mathcal{K}^*(\alpha^2\mathcal{K}\psi^o) \;-\; \lambda\,\psi^o \;=\; 0\,, \quad \|\psi^o\|_X = 1\,, \tag{7.7}$$

for the compact, self-adjoint and positive definite operator $\mathcal{K}^*\alpha^2\mathcal{K}$ in X. We observe that $\lambda \;=\; \|\alpha\mathcal{K}\psi^o\|^2_{L^2(S^{d-1})}$. We can summarize this particular but important case as follows:

[2] As an example we could take $g(\psi) = \|\psi\|^2_X - 1$ in which case U is simply the unit ball in X.

Theorem 7.2. *Let the assumptions (A1)–(A2) be satisfied. The maximum of* $\mathcal{J}(\psi) = \|\alpha\mathcal{K}\psi\|^2_{L^2(S^{d-1})}$ *on the set* $U = \{\psi \in X : \|\psi\|_X \leq 1\}$ *is equivalent to finding the maximal eigenvalue and a corresponding eigenvector of the compact, self-adjoint and positive definite operator* $\mathcal{K}^*\alpha^2\mathcal{K}$ *on* X*. More precisely, for every solution* $\psi^o \in U$ *of the optimization problem it is* $\|\psi^o\|_X = 1$ *and the optimal value* $\lambda = \|\alpha\mathcal{K}\psi^o\|^2_{L^2(S^{d-1})}$ *is the largest eigenvalue of* $\mathcal{K}^*\alpha^2\mathcal{K} : X \to X$ *and* ψ^o *a corresponding eigenvector. On the other hand, if* $\lambda > 0$ *is the largest eigenvalue of* $\mathcal{K}^*\alpha^2\mathcal{K}$ *and* ψ^o *a corresponding eigenvector, normalized to* $\|\psi^o\|_X = 1$*, then* ψ^o *solves the optimization problem with optimal value* λ*.*

Proof: In light of our previous discussion it remains to prove the second assertion. Let

$$\lambda_1 \geq \lambda_2 \geq \cdots > 0$$

be the ordered sequence of the eigenvalues of $\mathcal{K}^*\alpha^2\mathcal{K}$ with corresponding normalized eigenvectors $\psi_n \in X$, $n = 1, 2, \ldots$ Let $\psi \in X$ be arbitrary with $\|\psi\|_X = 1$. Then ψ has a representation in the form

$$\psi = \psi_0 + \sum_{n=1}^{\infty} (\psi, \psi_n)_X \, \psi_n$$

for some ψ_0 with $K^*(\alpha^2 K\psi_0) = 0$ and

$$\|\alpha\,\mathcal{K}\psi\|^2_{L^2(S^{d-1})} = \left(\mathcal{K}^*(\alpha^2\mathcal{K}\psi), \psi\right)_X = \sum_{n=1}^{\infty} \lambda_n \left|(\psi, \psi_n)_X\right|^2$$

$$\leq \lambda_1 \sum_{n=1}^{\infty} \left|(\psi, \psi_n)_X\right|^2 \leq \lambda_1 \, \|\psi\|^2_X = \lambda_1 \,.$$

Finally, the choice $\psi = \psi_1$ yields $\|\alpha\,\mathcal{K}\psi_1\|^2_{L^2(S^{d-1})} = \lambda_1$. □

This theorem shows that the question of uniqueness of an optimal solution is closely related to the multiplicity of the largest eigenvalue of $\mathcal{K}^*\alpha^2\mathcal{K}$. Since this operator is compact we know from the theorem of Riesz (see A.40) that there exist only finitely many linear independent eigenvectors $\psi_1^o, \ldots, \psi_N^o \in X$ corresponding to λ_1. Therefore, the set $\varPhi$ of all optimal solutions in this case is given by

$$\varPhi = \left\{ \left\| \sum_{j=1}^{N} \rho_j \psi_j^o \right\|_X^{-1} \sum_{j=1}^{N} \rho_j \psi_j^o : \rho_j \in \mathbb{C} \right\}.$$

For the numerical computations, one restricts the maximization of $\mathcal{J}$ to finite dimensional subspaces X_n of X, which can be accomplished by applying Theorem 3.25.

Theorem 7.3. *Let the assumptions (A1)–(A2) be satisfied. Assume, furthermore, that $X_n \subset X$ is an ultimately dense sequence of finite dimensional subspaces. Finding the maximum value of $\mathcal{J}(\psi) = \|\alpha\mathcal{K}\psi\|^2_{L^2(S^{d-1})}$ on the set $U_n = \{\psi \in X_n : \|\psi\|_X \leq 1\}$ and the optimal solution is equivalent to finding the maximal eigenvalue λ_n and a corresponding normalized eigenvector $\psi_n^o \in X_n$ of the self-adjoint and positive definite operator $P_n\mathcal{K}^*\alpha^2\mathcal{K}\big|_{X_n} : X_n \to X_n$ where $P_n : X \to X_n$ denotes the orthogonal projection onto X_n. Furthermore, any sequence $\psi_n^o \in U_n$ of optimal solutions has accumulation points and every such accumulation point is optimal for the maximization of $\mathcal{J}$ on $U = \{\psi \in X : \|\psi\|_X \leq 1\}$, i.e. an eigenfunction of $\mathcal{K}^*\alpha^2\mathcal{K}$ corresponding to the largest eigenvalue λ. Finally, λ_n converges to λ as $n \to \infty$.*

Proof: We set $\mathcal{K}_n := \mathcal{K}\big|_{X_n}$ and apply the previous theorem to X_n and $\mathcal{K}_n$ in place of X and $\mathcal{K}$, respectively. We note that $\mathcal{K}_n^* = P_n\mathcal{K}$ since

$$\left(\mathcal{K}_n\psi_n, \phi\right)_{L^2(S^{d-1})} = \left(\mathcal{K}\psi_n, \phi\right)_{L^2(S^{d-1})} = \left(\psi_n, \mathcal{K}^*\phi\right)_X = \left(\psi_n, P_n\mathcal{K}^*\phi\right)_X$$

for all $\psi_n \in X_n$ and $\phi \in L^2(S^{d-1})$. Application of Theorem 3.25 shows that any sequence $\psi_n^o \in U_n$ of optimal solutions has weak accumulation points, and every such weak accumulation point ψ^o is optimal. In particular, $\|\psi^o\|_X = 1$. It remains to show that every weak convergent sequence $\psi_n^o \in X_n$ of normalized optimal solutions is also convergent with respect to the norm. Let $\{\psi_n^o\}$ be weakly convergent to ψ^o. Then $\|\psi_n^o\|_X = \|\psi^o\|_X = 1$ and we conclude that

$$\begin{aligned} \|\psi_n^o - \psi^o\|_X^2 &= \|\psi_n^o\|_X^2 + \|\psi^o\|_X^2 - 2\,\mathrm{Re}\left(\psi_n^o, \psi^o\right)_X \\ &= 2\left[1 - \mathrm{Re}\left(\psi_n^o, \psi^o\right)_X\right] \longrightarrow 2\left[1 - \mathrm{Re}\left(\psi^o, \psi^o\right)_X\right] = 0 \end{aligned}$$

as n tends to infinity. □

Remark: We note that, in principle, it makes no difference if we first discretize the problem, i.e. restrict the objective function to the finite dimensional subspace, and then apply the multiplier rule to the finite dimensional system or apply, first, the multiplier rule to the infinite dimensional system and then use the projection method to solve the finite dimensional eigenvalue problem.

In the remaining part of this subsection we study the maximization of the power intensities in given directions $\hat{\boldsymbol{x}}_j \in S^{d-1}$, $j = 1, \ldots, m$, on constraint sets U of the form (7.5). In this case, the performance functional is given by (7.3), i.e.

$$\mathcal{J}(\psi) := \sum_{j=1}^{m} w_j \left|(\mathcal{K}\psi)(\hat{\boldsymbol{x}}_j)\right|^2, \quad \psi \in X, \tag{7.8}$$

where $w_j > 0$ are given weights and $\mathcal{K} : X \to C(S^{d-1})$ satisfies the condition (A1), i.e. is compact and one-to-one. We note that this case can formally be subsumed under the previous one by defining α as a sum of delta-functions:

$$\alpha(\hat{\boldsymbol{x}}) \ := \ \sum_{j=1}^{m} \sqrt{w_j}\, \delta(|\hat{\boldsymbol{x}} - \hat{\boldsymbol{x}}_j|)\,.$$

By Lemma 7.1 the functional $\mathcal{J}$ is still convex (but no longer strictly convex) and is differentiable. Its Fréchet derivative is

$$\nabla \mathcal{J}(\psi) \ = \ 2 \sum_{j=1}^{m} w_j\, (\mathcal{K}\psi)(\hat{\boldsymbol{x}}_j)\, p_j$$

where the $p_j \in X$ are defined by the Riesz representation of the bounded linear functional $\varphi \mapsto (\mathcal{K}\varphi)(\hat{\boldsymbol{x}}_j)$, $\varphi \in X$, i.e. for each $j = 1, 2, \ldots, m$, the element $p_j \in X$ is the unique element of X for which

$$(\mathcal{K}\varphi)(\hat{\boldsymbol{x}}_j) \ = \ \bigl(\varphi, p_j\bigr)_X \quad \text{for all } \varphi \in X\,. \tag{7.9}$$

By Theorem 3.16 there exist solutions of the optimization problem

$$\text{Maximize} \quad \sum_{j=1}^{m} w_j\ |(\mathcal{K}\psi)(\hat{\boldsymbol{x}}_j)|^2 \quad \text{subject to } \psi \in U\,, \tag{7.10}$$

and at least one solution is attained at an extreme point of U. Application of the Lagrange multiplier rule (Theorem 3.22) yields:

Theorem 7.4. *Let $\hat{\boldsymbol{x}}_j \in S^{d-1}$, $j = 1, \ldots, m$, and $\mathcal{K} : X \to C(S^{d-1})$ be compact and one-to-one. Furthermore, let $g : X \to \mathbb{R}$ be uniformly convex and continuously differentiable. Then there exist optimal solutions of (7.10) where $U = \{\psi \in X : g(\psi) \leq 0\}$. Furthermore, for any optimal solution ψ^o with $\nabla g(\psi^o) \neq 0$ of this optimization problem there exists $\lambda \in \mathbb{R}$, $\lambda \geq 0$, with*

$$\sum_{j=1}^{m} w_j\, (\mathcal{K}\psi^o)(\hat{\boldsymbol{x}}_j)\, p_j \ - \ \lambda \nabla g(\psi^o) \ = \ 0\,. \tag{7.11}$$

Again, the $p_j \in X$ are given by the Riesz representation of the mapping $\varphi \mapsto (\mathcal{K}\varphi)(\hat{\boldsymbol{x}}_j)$, $\varphi \in X$. If the optimal value $\sum_{j=1}^{m} w_j\ |(\mathcal{K}\psi^o)(\hat{\boldsymbol{x}}_j)|^2 > 0$ the multiplier λ is strictly positive and $g(\psi^o) = 0$.

Proof (of the last assertion): We know from the multiplier rule that $\lambda\, g(\psi^o) = 0$, i.e. λ or $g(\psi^o)$ vanishes - or both. Assume that $\lambda = 0$. Then (7.11) yields

$$\sum_{j=1}^{m} w_j\, (\mathcal{K}\psi^o)(\hat{\boldsymbol{x}}_j)\, p_j \ = \ 0\,.$$

Multiplication of this equation by ψ^o and using (7.9) leads to
$\sum_{j=1}^{m} w_j\ |(\mathcal{K}\psi^o)(\hat{\boldsymbol{x}}_j)|^2 = 0$ which contradicts the assumption. Therefore, $\lambda > 0$ and thus also $g(\psi^o) = 0$. □

In the particular case that U is the unit ball in X we have that $\nabla g(\psi) = 2\,\psi$ and the multiplier rule (7.11) simplifies to

$$\sum_{j=1}^{m} w_j\,(\mathcal{K}\psi^o)(\hat{\boldsymbol{x}}_j)\,p_j \;=\; \lambda\,\psi^o\,. \tag{7.12}$$

Therefore, ψ^o is a linear combination of the p_j, $j = 1, \ldots, m$. If we make the ansatz

$$\psi^o \;=\; \sum_{j=1}^{m} a_j\,\sqrt{w_j}\,p_j$$

for some $a_j \in \mathbb{C}$, equation (7.12) leads to the eigenvalue problem $Ma = \lambda a$ for the Hermitian matrix $M \in \mathbb{C}^{m\times m}$ given by

$$M_{ij} \;=\; \sqrt{w_i\,w_j}\,(\mathcal{K}p_j)(\hat{\boldsymbol{x}}_i) \;=\; \sqrt{w_i\,w_j}\,(p_j, p_i)_X\,, \quad i,j = 1, \ldots, m\,. \tag{7.13}$$

Multiplication of (7.12) by ψ^o then leads to $\lambda = \sum_{j=1}^{m} w_j\,|(\mathcal{K}\psi^o)(\hat{\boldsymbol{x}}_j)|^2$. We have therefore to determine the largest eigenvalue of M and the corresponding eigenvector $a \in \mathbb{C}^m$, normalized such that $\left\|\sum_{j=1}^{m} a_j\,\sqrt{w_j}\,p_j\right\|_X = 1$.
Computations for this particular constrained problem are completely analogous to the previous computations. Instead of repeating that analysis in concrete examples e.g., for the line source, we turn to the consideration of problems in which the constraints themselves are described by pointwise conditions. We analyze the general situation first, and then illustrate by applying the results to the line source in §7.2.3.

7.2.2 Pointwise Constraints on Inputs

For this analysis we take the Hilbert space $X = L^2(\Gamma)$ for some curve or surface $\Gamma \subset \mathbb{R}^d$, and restrict ourselves to the case where $L^2(\Gamma)$ consists of complex-valued but scalar functions. The case of vector fields as occurs in the study of boundary controls for Maxwell's equations can be treated in a similar fashion. We consider the power optimization problem with cost functional $\mathcal{J}(\psi) = \|\alpha\mathcal{K}\psi\|^2_{L^2(S^{d-1})}$ only, i.e.,

$$\text{Maximize} \quad \mathcal{J}(\psi) \;=\; \|\alpha\,\mathcal{K}\psi\|^2_{L^2(S^{d-1})} \quad \text{subject to } \psi \in U\,, \tag{7.14}$$

where the the set U of constraints is given as in (3.10) by

$$U \;=\; \left\{\psi \in L^2(\Gamma) : \psi(\boldsymbol{x}) \in V(\boldsymbol{x}) \text{ a.e. on } \Gamma\right\},$$

for $V(\boldsymbol{x}) \subset \mathbb{C}$ closed and convex for each $\boldsymbol{x}$, and the set $\bigcup_{\boldsymbol{x}\in\Gamma} V(\boldsymbol{x})$ bounded with the graph of V measurable. We have seen that optimal solutions ψ^o exist and are necessarily extreme points since the functional $\mathcal{J}$ is strictly convex. Therefore, $\psi^o(\boldsymbol{x}) \in \operatorname{ext} V(\boldsymbol{x})$ for almost all $\boldsymbol{x} \in \Gamma$. Application of Lemma 3.21 and Lemma 7.1 yields

$$\operatorname{Re} \int_\Gamma \overline{v(\boldsymbol{x})}\bigl[\varphi(\boldsymbol{x}) - \psi^o(\boldsymbol{x})\bigr]\, ds(\boldsymbol{x}) \;\le\; 0 \quad \text{for all } \varphi \text{ with } \varphi(\boldsymbol{x}) \in V(\boldsymbol{x}) \quad (7.15)$$

for almost all $\boldsymbol{x} \in \Gamma$ where we have set $v := \mathcal{K}^*(\alpha^2 \mathcal{K}\psi^o) \in L^2(\Gamma)$. From this inequality we may conclude that the optimal solution satisfies the pointwise inequality

$$\operatorname{Re}\bigl\{\overline{v(\boldsymbol{x})}\bigl[z - \psi^o(\boldsymbol{x})\bigr]\bigr\} \;\le\; 0 \quad \text{for all } z \in V(\boldsymbol{x}) \text{ and almost all } \boldsymbol{x} \in \Gamma\,. \quad (7.16)$$

Indeed, suppose that (7.16) fails to hold. Then there is a set $I \subset \Gamma$ of positive measure such that the sets

$$W(\boldsymbol{x}) \;:=\; \Bigl\{z \in V(\boldsymbol{x}) : \operatorname{Re}\bigl\{\,\overline{v(\boldsymbol{x})}\bigl[z - \psi^o(\boldsymbol{x})\bigr]\,\bigr\} > 0\Bigr\}$$

are non-empty for all $\boldsymbol{x} \in I$. The graph of W is measurable. Therefore, by a measurable selection theorem (see the remarks following Definition 3.11, or [22]), there exists a measurable function $\varphi : I \to \mathbb{C}$ with $\varphi(\boldsymbol{x}) \in W(\boldsymbol{x})$ a.e. on I. We extend φ by setting $\varphi(\boldsymbol{x}) = \psi^o(\boldsymbol{x})$ for $\boldsymbol{x} \in \Gamma \setminus I$. Then $\varphi \in L^2(\Gamma)$ and $\varphi(\boldsymbol{x}) \in V(\boldsymbol{x})$ a.e. and $\operatorname{Re}\bigl\{\overline{v(\boldsymbol{x})}\bigl[\varphi(\boldsymbol{x}) - \psi^o(\boldsymbol{x})\bigr]\bigr\} \ge 0$ on Γ and is strictly positive on a set of positive measure. Integration yields $\operatorname{Re}\bigl(v, \varphi - \psi^o\bigr)_{L^2(\Gamma)} > 0$ which contradicts (7.15).

We point out that the variational inequality (7.16) not only yields the extremal property of $\psi^o(\boldsymbol{x})$ for almost all $\boldsymbol{x} \in \Gamma$ but, in some cases, even more. Let us consider the special case, where $V = V(\boldsymbol{x})$ is constant with respect to $\boldsymbol{x}$ and, in particular, is either a rectangle or a disc in $\mathbb{C}$ and that the set $\Gamma \subset \mathbb{R}^2$ is the analytic boundary of some plane region $\Omega \subset \mathbb{R}^2$ with connected exterior. Then, $\psi^o(\boldsymbol{x})$ is an extreme point of V for almost all $\boldsymbol{x} \in \Gamma$. We will need a further assumption on the operator $\mathcal{K} : L^2(\Gamma) \to C(S^{d-1})$ which we formulate in terms of its L^2-adjoint $\mathcal{K}^* : L^2(S^{d-1}) \to L^2(\Gamma)$ as:

(A4) $\mathcal{K}^*\varphi$ is analytic on the analytic curve $\Gamma \subset \mathbb{R}^2$ for every $\varphi \in L^2(S^{d-1})$.

The assumption (A4) and the variational formula (7.16) now yield a **finite bang-bang principle**:

Theorem 7.5. *Let ψ^o be a solution of (7.14) and let the assumptions (A1)–(A4) hold (for $d = 2$). Again, set $v := \mathcal{K}^*(\alpha^2 \mathcal{K}\psi^o)$ on Γ.*

(a) If $V \subset \mathbb{C}$ is a disc with center 0 and radius R then ψ^o coincides a.e. with the piecewise continuous function

$$\hat{\psi}(\boldsymbol{x}) \;=\; R \operatorname{sign} v(\boldsymbol{x})\,,$$

where $\operatorname{sign} z = z/|z|$ denotes the sign of $z \in \mathbb{C}$, $z \neq 0$.

(b) If $V \subset \mathbb{C}$ is the rectangle $V = [a_-, a_+] + i[b_-, b_+]$ then ψ^o coincides a.e. with the piecewise continuous function $\hat{\psi}$ given by

$$\operatorname{Re}\hat{\psi}(\boldsymbol{x}) \;=\; \begin{cases} a_-\,, & \operatorname{Re} v(\boldsymbol{x}) > 0\,, \\ a_+\,, & \operatorname{Re} v(\boldsymbol{x}) < 0\,, \end{cases} \qquad \operatorname{Im}\hat{\psi}(\boldsymbol{x}) \;=\; \begin{cases} b_-\,, & \operatorname{Im} v(\boldsymbol{x}) > 0\,, \\ b_+\,, & \operatorname{Im} v(\boldsymbol{x}) < 0\,. \end{cases}$$

Proof: (a) Substituting $z = R \operatorname{sign} v(\boldsymbol{x})$ into equation (7.16) yields

$$R\,|v(\boldsymbol{x})| \;\leq\; \operatorname{Re}\left\{\overline{v(\boldsymbol{x})}\,\psi^o(\boldsymbol{x})\right\} \quad \text{for almost all } \boldsymbol{x} \in \Gamma .$$

From this we conclude

$$\begin{aligned} 0 \;\leq\; \left|\psi^o(\boldsymbol{x}) - R \operatorname{sign} v(\boldsymbol{x})\right|^2 &= |\psi^o(\boldsymbol{x})|^2 \;+\; R^2 \;-\; \frac{2R}{|v(\boldsymbol{x})|} \operatorname{Re}\left\{\overline{v(\boldsymbol{x})}\,\psi^o(\boldsymbol{x})\right\} \\ &\leq 2R^2 \;-\; 2R^2 \;=\; 0 \end{aligned}$$

almost everywhere.
(b) Substituting $z = t + i \operatorname{Im} \psi^o(\boldsymbol{x})$ and $z = \operatorname{Re} \psi^o(\boldsymbol{x}) + is$ into equation (7.16) yields

$$\operatorname{Re} v(\boldsymbol{x})\,\left[t - \operatorname{Re} \psi^o(\boldsymbol{x})\right] \leq 0 \quad \text{and} \quad \operatorname{Im} v(\boldsymbol{x})\,\left[s - \operatorname{Im} \psi^o(\boldsymbol{x})\right] \leq 0$$

for all $t \in [a_-, a_+]$ and $s \in [b_-, b_+]$ from which the assertion follows. □

7.2.3 Numerical Simulations

In this subsection we study optimization problems for two particular cases and present numerical results.

As a first example we consider a **circular line source**. In the plane of the line source of radius a the operator K takes the form (see Subsection 1.5.2)

$$(K\psi)(\phi) \;=\; \int_0^{2\pi} \psi(s)\, \mathrm{e}^{-ika\cos(\phi - s)}\, ds\,, \quad 0 \leq \phi \leq 2\pi\,.$$

Then we consider the problem:

$$\begin{aligned} &\text{Maximize} && \int_0^{2\pi} \alpha(\phi)\, \left|(K\psi)(\phi)\right|^2 d\phi \\ &\text{subject to} && \psi \in L^2(0, 2\pi)\,, \quad \int_0^{2\pi} |\psi(t)|^2\, dt \;\leq\; 1\,. \end{aligned}$$

For the finite dimensional approximations we restrict ψ to lie in

$$X_n \;=\; \operatorname{span}\left\{\mathrm{e}^{ijt} : |j| \leq n\right\}.$$

Using the Jacobi-Anger expansion (1.14) we can represent $K\psi$ as a Fourier series in the form

$$(K\psi)(\phi) \;=\; 2\pi \sum_{m \in \mathbb{Z}} \psi_m\, (-i)^m\, J_m(ka)\, \mathrm{e}^{im\phi}\,, \quad 0 \leq \phi \leq 2\pi\,,$$

where

$$\psi_m \;=\; \frac{1}{2\pi}\int\limits_0^{2\pi} \psi(t)\,\mathrm{e}^{-imt}\,dt\,, \quad m\in\mathbb{Z}\,,$$

are the Fourier coefficients of $\psi \in L^2(0,2\pi)$.

Example 7.6. As a particular case we take α to the characteristic function of the interval $[\alpha_1,\alpha_2]\subset[0,2\pi]$ where $\alpha_1<\alpha_2$. Then the optimization problem has the form:

$$\begin{aligned} &\text{Maximize} \quad \int\limits_{\alpha_1}^{\alpha_2} \left|(K\psi)(\phi)\right|^2 d\phi \\ &\text{subject to} \quad \psi\in L^2(0,2\pi)\,, \quad \int\limits_0^{2\pi} |\psi(t)|^2\,dt \;\le\; 1\,. \end{aligned}$$

For the application of Theorem 7.3 we have to compute the operator $P_nK^*\alpha^2K$. Analogously to K the operator $K^*\alpha^2K$ is given by

$$(K^*\alpha^2K\psi)(t) \;=\; 2\pi\sum_{\ell,m\in\mathbb{Z}} \psi_m\, i^{m-\ell}\, J_m(ka)\, J_\ell(ka)\, \frac{\mathrm{e}^{i(m-\ell)\alpha_2}-\mathrm{e}^{i(m-\ell)\alpha_1}}{i\,(m-\ell)}\,\mathrm{e}^{i\ell t}$$

(in the case $\ell=m$ the fraction has to be replaced by $\alpha_2-\alpha_1$). Therefore, the finite dimensional operator $P_nK^*\alpha^2K\big|_{X_n}$ is represented by the matrix $A=\left(a_{\ell m}\right)\in\mathbb{C}^{(2n+1)\times(2n+1)}$ where

$$a_{\ell m} \;=\; \begin{cases} 2\pi\, i^{m-\ell}\, J_m(ka)\, J_\ell(ka)\, \dfrac{\mathrm{e}^{i(m-\ell)\alpha_2}-\mathrm{e}^{i(m-\ell)\alpha_1}}{i\,(m-\ell)}\,, & \ell\neq m\,, \\ 2\pi\, J_m(ka)^2\,(\alpha_2-\alpha_1)\,, & \ell=m\,. \end{cases}$$

The plots of Figure 7.1 show $|K\psi^o|$ of the numerical calculations for the angular sector $[0,\pi/4]$ and wave lengths $\lambda=1$ and $\lambda=\pi$, respectively.

Having treated the circular line source we now turn the case of a **linear line source** of length 2ℓ along the $\hat{e}_3$-axis and polarization vector $\hat{\boldsymbol{p}}=\hat{e}_3$. We have seen in Sections 1.5 and 4.5 that the electric far field pattern for a linear line source can be written as (see (1.49))

$$|\boldsymbol{E}_\infty(\theta)| \;=\; \frac{\omega\mu_0}{4\pi}\,\sin\theta\left|\int\limits_{-\ell}^{\ell}\psi(s)\,\mathrm{e}^{-iks\cos\theta}\,ds\right|\,, \quad \theta\in[0,\pi]\,, \tag{7.17}$$

in which case the far field operator $K: L^2(-\ell,\ell)\to C[-1,+1]$ is

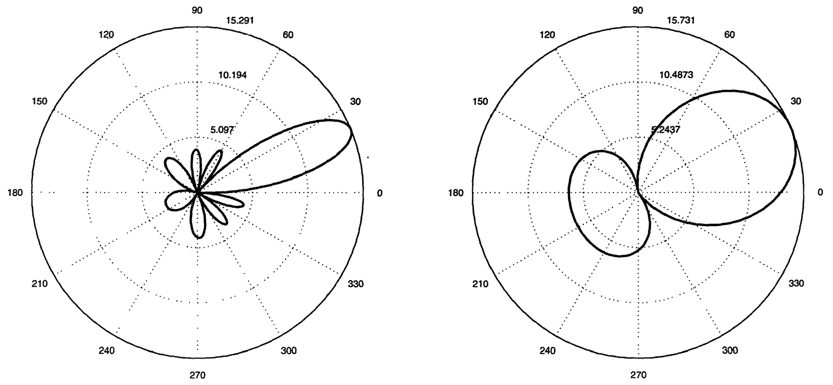


Fig. 7.1. Plots of $|\mathcal{K}\psi^o|$ for wave lengths $\lambda = 1$ and $\lambda = \pi$

$$(K\psi)(t) \;=\; \sqrt{1-t^2}\int\limits_{-\ell}^{\ell} \psi(s)\,\mathrm{e}^{-ikts}\,ds\,, \quad |t| \le 1\,, \tag{7.18}$$

where we have made the substitution $t = \cos\theta$. In particular, we note that the far field is independent of the angular variable $\phi \in [0, 2\pi]$. The adjoint of the far field operator is then given by

$$(K^*\varphi)(t) \;=\; \int\limits_{-1}^{1} \sqrt{1-s^2}\,\varphi(s)\,\mathrm{e}^{ikts}\,ds\,, \quad |t| \le \ell\,. \tag{7.19}$$

Again let $\mathcal{A} \subset S^{d-1}$ be some subset of the sphere which is open relative to S^{d-1}. We study the following optimization problems corresponding to (7.1) and (7.3) for $U = \{\psi \in X : \|\psi\|_X \le 1\}$:

$$\begin{aligned} &\text{Maximize} \quad \|\boldsymbol{E}_\infty\|_{L^2(\mathcal{A})} \;=\; \int\limits_{\mathcal{A}} |(K\psi)(\cos\theta)|^2 ds \qquad &(7.20\text{a})\\ &\text{subject to} \quad \|\psi\|^2_{L^2(-\ell,\ell)} \;\le\; 1\,, \end{aligned}$$

and

$$\begin{aligned} &\text{maximize} \quad \sum_{j=1}^{m} w_j\,|\boldsymbol{E}_\infty(\hat{\boldsymbol{x}}_j)|^2 \;=\; \sum_{j=1}^{m} w_j\,|(K\psi)(\cos\theta_j)|^2 \qquad &(7.20\text{b})\\ &\text{subject to} \quad \|\psi\|^2_{L^2(-\ell,\ell)} \;\le\; 1\,, \end{aligned}$$

where $\hat{\boldsymbol{x}}_j \in S^{d-1}$ and $w_j > 0$ are given and $(\theta_j, \phi_j) \in [0,\pi] \times [0, 2\pi)$ are the polar coordinates of $\hat{\boldsymbol{x}}_j$, $j = 1, \ldots, m$.

Let us first study (7.20a). For this, let α be the characteristic function of the patch $\mathcal{A}$, parametrized in spherical polar coordinates (θ, ϕ). We compute

$$\int_{\mathcal{A}} |(K\psi)(\cos\theta)|^2 ds = \int_0^{\pi}\int_0^{2\pi} \alpha(\theta,\phi)\, d\phi\, |(K\psi)(\cos\theta)|^2 \sin\theta\, d\theta$$
$$= \int_{-1}^{1} \tilde{\alpha}(t)^2\, |(K\psi)(t)|^2\, dt \;=\; \|\tilde{\alpha}\, K\psi\|^2_{L^2(-1,+1)}\,,$$

where $\tilde{\alpha}(t)^2 = \int_0^{2\pi} \alpha(\arccos t, \phi)\, d\phi$, $|t| \le 1$. The total power corresponds to $\alpha \equiv 1$ i.e., $\tilde{\alpha} \equiv \sqrt{2\pi}$.

The maximization problem (7.20a) is therefore equivalent to:

$$\text{Maximize} \quad \|\tilde{\alpha}\, K\psi\|^2_{L^2(-1,1)} \quad \text{subject to} \quad \|\psi\|_{L^2(-\ell,\ell)} \;\le\; 1\,. \tag{7.21}$$

We want to apply Theorem 7.2 to $\mathcal{K} = K : L^2(-\ell,\ell) \to C[-1,1]$ and to do so we must compute the operator $K^*(\tilde{\alpha}^2 K)$. Using the form of the adjoint K^* (7.19) we see easily by interchanging the orders of integration that

$$(K^*\tilde{\alpha}^2 K\psi)(t) \;=\; \int_{-\ell}^{\ell} \psi(s)\, A(t-s)\, ds\,, \quad |t| \le \ell\,, \tag{7.22}$$

with kernel

$$A(\tau) \;=\; \int_{-1}^{1} (1-s^2)\, \tilde{\alpha}(s)^2\, e^{iks\tau}\, ds\,, \quad \tau \in \mathbb{R}\,. \tag{7.23}$$

As remarked before, the case $\alpha \equiv 1$ corresponds to $\tilde{\alpha}^2 \equiv 2\pi$ and

$$(K^*K\psi)(t) \;=\; \int_{-\ell}^{\ell} \psi(s)\, a(t-s)\, ds\,, \quad |t| \le \ell\,, \tag{7.24}$$

where the kernel was formed, already in Section 4.5, to have the form

$$a(\tau) \;=\; 2\pi \int_{-1}^{1} (1-s^2)\, e^{iks\tau}\, ds \;=\; \frac{8\pi}{(k\tau)^2}\left[\frac{\sin(k\tau)}{k\tau} \;-\; \cos(k\tau)\right]\,, \quad \tau \in \mathbb{R}\,. \tag{7.25}$$

Therefore, we must find the largest eigenvalue λ and a corresponding normalized eigenfunction ψ of the eigenvalue problem

$$\lambda\, \psi(t) \;=\; \int_{-1}^{+1} A(t-s)\, \psi(s)\, ds\,, \quad |t| \le 1\,. \tag{7.26}$$

For the numerical computation we use the Nyström method and replace the integral by a quadrature rule of the form

$$\int_{-1}^{+1} f(s)\,ds \;\approx\; \sum_{j=1}^{n} q_j\, f(s_j) \tag{7.27}$$

where s_j and q_j are the Gauss-Legendre nodes and weights, respectively. Substituting $t = s_i$ in (7.26) leads to the approximate equation

$$\lambda\,\psi^{(n)}(s_i) \;=\; \sum_{j=1}^{n} q_j\, A(s_i - s_j)\,\psi^{(n)}(s_j)\,, \quad i = 1,\dots,n\,. \tag{7.28}$$

Since the matrix of this equation is not Hermitian we symmetrize by multiplying the equation by $\sqrt{q_i}$ and setting $\psi_j := \sqrt{q_j}\,\psi^{(n)}(s_j)$, $j = 1,\dots,n$. The eigenvalue problem then becomes

$$\lambda\,\psi_i \;=\; \sum_{j=1}^{n} \sqrt{q_j\,q_i}\,A(s_i - s_j)\,\psi_j\,, \quad i = 1,\dots,n\,, \tag{7.29}$$

for the Hermitian matrix $M_{ij} = \sqrt{q_j\,q_i}\,A(s_i - s_j)$, $i, j = 1,\dots,n$.

Before we present actual numerical results we consider the second of the optimization problems, (7.20b). We can assume without loss of generality that $\theta_j \in (0,\pi)$ since otherwise $(K\psi)(\cos\theta_j) = 0$ and the corresponding term would not appear in the cost functional.

In order to apply the Lagrange multiplier rule of Theorem 7.4 we need to compute the Riesz representation of the functional $\psi \mapsto (K\psi)(\cos\theta_j)$, $\psi \in L^2(-\ell,\ell)$, $j = 1,\dots,m$. But this is obvious since

$$(K\psi)(\cos\theta_j) \;=\; \sin\theta_j \int_{-\ell}^{\ell} \psi(s)\,\mathrm{e}^{-iks\cos\theta_j}\,ds \;=\; (\psi, p_j)_{L^2(-\ell,\ell)} \tag{7.30}$$

with $p_j(s) = \sin\theta_j \exp(iks\cos\theta_j)$, $s \in (-\ell,\ell)$. For pairwise different θ_j these functions are linearly independent, and therefore, we can apply Theorem 7.4. In particular, $\|\psi^o\|_{L^2(-\ell,\ell)} = 1$, and the optimal function ψ^o has the form

$$\psi^o(s) \;=\; \sum_{j=1}^{m} a_j\,\sqrt{w_j}\,p_j(s) \;=\; \sum_{j=1}^{m} a_j\,\sqrt{w_j}\,\sin\theta_j\,\mathrm{e}^{iks\cos\theta_j} \quad \text{for some } a_j \in \mathbb{C}$$

where $\boldsymbol{a} = (a_j) \in \mathbb{C}^m$ is an eigenvector of the matrix $M \in \mathbb{C}^{m\times m}$ given by

$$\begin{aligned} M_{\nu\mu} = \sqrt{w_\nu\,w_\mu}\,\bigl(p_\nu, p_\mu\bigr)_{L^2(-\ell,\ell)} \;&=\; \sqrt{w_\nu\,w_\mu}\,\sin\theta_\nu\,\sin\theta_\mu \int_{-\ell}^{\ell} \mathrm{e}^{iks(\cos\theta_\nu - \cos\theta_\mu)}\,ds \\ = 2\ell\,\sqrt{w_\nu\,w_\mu}\,\sin\theta_\nu\,\sin\theta_\mu\,&\frac{\sin\bigl[k\ell(\cos\theta_\nu - \cos\theta_\mu)\bigr]}{k\ell(\cos\theta_\nu - \cos\theta_\mu)} \end{aligned} \tag{7.31}$$

which corresponds to the largest eigenvalue λ_{max} and is normalized to $|a|_2 = 1/\lambda_{max}$. $K\psi^o$ is given by

$$(K\psi^o)(t) \;=\; 2\ell\sqrt{1-t^2}\sum_{j=1}^{m} a_j\sqrt{w_j}\sin\theta_j\frac{\sin\bigl[k\ell(\cos\theta_j - t)\bigr]}{k\ell(\cos\theta_j - t)}\,, \quad |t| \leq 1\,.$$

Example 7.7. We now make some particular choices in the case of the linear line source. We compare the optimization problems (7.20a) and (7.20b) for the same values of θ_1, θ_2 and for a fixed wave number k.

Specifically, we consider (7.20a) for the case where $\mathcal{A}$ consists of two separate strips, i.e.,

$$\alpha(\theta,\phi) \;:=\; \begin{cases} c_j, \; |\theta - \theta_j| \;\leq\; \delta_j, & j = 1 \text{ or } 2, \\ 0, \;\; \text{otherwise}, \end{cases} \tag{7.32}$$

and $c_j > 0$ is chosen such that the areas of the corresponding strips $\mathcal{A}_j$, $j = 1, 2$, are equal. This leads to $c_j = 1/(\sin\theta_j\sin\delta_j)$, $j = 1, 2$. Then we have that

$$\tilde{\alpha}(t)^2 \;=\; 2\pi\begin{cases} c_j\,, \; \bigl|t - \underbrace{\cos\theta_j\cos\delta_j}_{=:t_j}\bigr| \;\leq\; \underbrace{\sin\theta_j\sin\delta_j}_{=:\Delta_j}, & j = 1 \text{ or } 2, \\ 0\,, \text{ otherwise}\,. \end{cases}$$

For this case the kernel A from (7.23) takes the specific form

$$A(\tau) \;=\; \int_{-1}^{1}(1-s^2)\,\tilde{\alpha}(s)^2\,e^{iks\tau}\,ds \;=\; 2\pi\sum_{j=1}^{2} c_j\int_{t_j-\Delta_j}^{t_j+\Delta_j}(1-s^2)\,e^{iks\tau}\,ds$$

and can be computed explicitly. We denote the optimal value for this problem by ψ_a^o.

Turning to the study of the optimization problem (7.20b) for $m = 2$ and $w_j = 1$, $j = 1, 2$ we use the same values of θ_j and k. Then the matrix M from (7.31) takes the form

$$M \;=\; 2\ell\begin{pmatrix} \sin^2\theta_1 & S \\ S & \sin^2\theta_2 \end{pmatrix} \quad \text{with} \quad S \;=\; \sin\theta_1\sin\theta_2\frac{\sin\bigl[k\ell(\cos\theta_1 - \cos\theta_2)\bigr]}{k\ell(\cos\theta_1 - \cos\theta_2)}\,.$$

The eigenvalues are given by

$$\ell\,(\sin^2\theta_1 + \sin^2\theta_2) \;\pm\; \ell\sqrt{4S^2 + (\sin^2\theta_1 - \sin^2\theta_2)^2}\,,$$

and the larger of these, λ_{max}, is the one with the plus sign. If $\boldsymbol{a} \in \mathbb{R}^2$ is the eigenvector corresponding to λ_{max}, normalized to $|a|_2 = 1/\lambda_{max}$, then the optimal solution ψ_b^o of (7.20b) is given by

$$\psi_b^o(s) \;=\; \sum_{j=1}^{2} a_j\sin\theta_j\exp(iks\cos\theta_j)\,, \; |s| \leq \ell\,,$$

and

$$(K\psi_b^o)(t) \;=\; 2\ell\sqrt{1-t^2}\sum_{j=1}^{2} a_j \sin\theta_j \frac{\sin\big[k\ell(\cos\theta_j - t)\big]}{k\ell(\cos\theta_j - t)}\,, \quad |t| \le 1\,.$$

For the following numerical computations we take the wave number $k = 10\pi$ (i.e. wave length $2\pi/k = 0.2$), the length of the line source $\ell = 1$, the width of the strips $\delta_1 = \delta_2 = 10^o$, and several values of θ_1 and θ_2. We use the notation

$$\mathcal{J}_a(\psi) \;=\; \|\tilde{\alpha} K\psi\|^2_{L^2(S_1)} \quad \text{and} \quad \mathcal{J}_b(\psi) \;=\; \sum_{j=1}^{2} |(K\psi_a^o)(\cos\theta_j)|^2\,.$$

	$\mathcal{J}_a(\psi_a^o)$	$\mathcal{J}_a(\psi_b^o)$	$\mathcal{J}_b(\psi_a^o)$	$\mathcal{J}_b(\psi_b^o)$
$\theta_1 = 30^o$ $\theta_2 = 150^o$	16.9	10.6	0.29	0.51
$\theta_1 = 31^o$ $\theta_2 = 150^o$	17.7	11.1	0.30	0.53
$\theta_1 = 50^o$ $\theta_2 = 150^o$	36.4	25.8	0.37	1.17

Plots of $|K\psi_a^o|$ and $|K\psi_b^o|$ for these examples are shown in Figure 7.2 – 7.4.

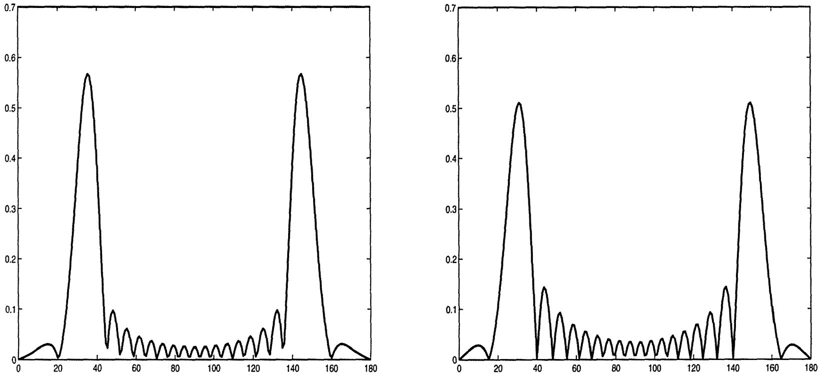


Fig. 7.2. $|K\psi_a^o|$ and $|K\psi_b^o|$ for $\theta_1 = 30^o$, $\theta_2 = 150^o$

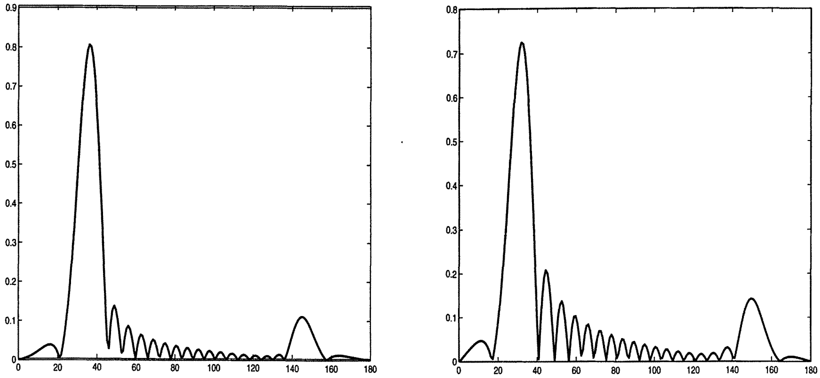


Fig. 7.3. $|K\psi_a^o|$ and $|K\psi_b^o|$ for $\theta_1 = 31^o$, $\theta_2 = 150^o$

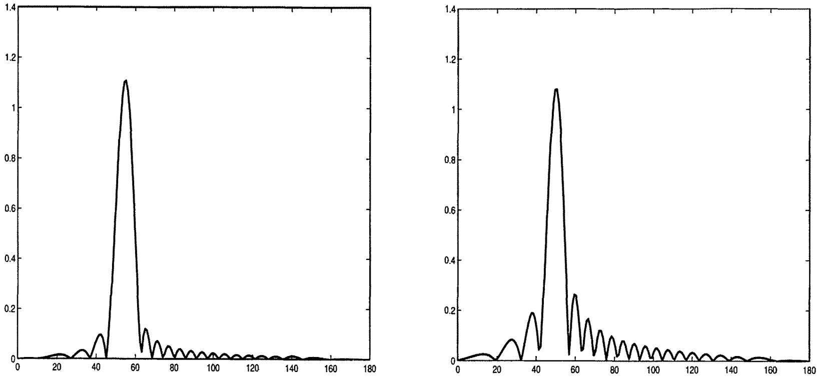


Fig. 7.4. $|K\psi_a^o|$ and $|K\psi_b^o|$ for $\theta_1 = 50^o$, $\theta_2 = 150^o$

We note that only in the case where θ_1 and θ_2 are symmetric with respect to $\pi/2$ we observe two beams. In all other cases the optimal pattern shows only one beam which is centered around the angle closest to $\pi/2$.

7.3 The Null-Placement Problem

In the previous discussions of the maximization of power in a sector of the far field, or of the directivity in a prescribed direction, little attention was paid to the question how the radiation pattern is affected by the designer's attempt to optimize the cost functional. We have mentioned, however, that it is possible to achieve highly directive patterns but at the cost of high (and undesirable) side lobe patterns. For the case of arrays, this was precisely the problem studied by Dolph [34] which we described in Chapter 1. We will return to a different approach to that problem again in Chapter 8.

Likewise, in the synthesis problem, far field patterns may be matched to within any desired degree of error, but only at the sacrifice of directivity or gain as was discussed by Taylor [133] and described previously in Chapter 4. An interesting analysis of the cost to the antenna design of producing a highly focused main beam may be found in the recent paper of Margetis et al. [89].

Yet another important class of problems are those for which we desire to maintain, or even optimize, the gain in a specific direction while controlling the far field pattern only in certain other preassigned directions. There are often sources of radiation, environmental or artificial, which come from a particular direction and interfere with the ability of the antenna to maintain its desirable performance. Often this interference appears after a desirable far field pattern has been established which is "efficient" as measured by a particular cost functional $\mathcal{J}$. What is wanted is to change the feedings to the antenna in such a way that the side lobes in the direction of the interfering signal are very low (the "placement of nulls") while maintaining, as closely as possible, the main beam characteristics e.g., the maximum power over the main beam sector, or the beam width.

Problems of this type arise, in practice, in communication problems where it is desirable to maintain gain in the direction of a remote antenna while reducing either localized interference due to background noise or to jamming originating from other known directions. Similar problems arise in radio astronomy.

Such problems are likewise optimization problems with constraints; but here the restrictions are to be imposed only over certain sectors of the far field, thereby allowing the radiation pattern in other directions to behave as necessary in order to contribute to the desired main beam performance. The particular mathematical formulation depends on the specific form of the cost functional to be optimized and on how the constraints are modeled. We discuss some typical models here to illustrate the flexibility of the general approach we have developed. In particular, we will consider two types of problems. It will be clear that similar problems may addressed with the techniques presented in this section.

In the first type of problem, we wish to maximize some functional related to the radiated power in some prescribed directions while keeping the radiated power small in some other, different, directions.

In the second problem we wish to modify an existing (and ostensibly desirable) far field pattern to reduce side lobes at prescribed locations or in specified sectors of the far field, while preserving as closely as possible the existing pattern outside these given positions.

As before, we assume that the operator $\mathcal{K} : X \longrightarrow C(S^{d-1})$ is a compact operator which satisfies the conditions (A1) and (A2) from the beginning of this chapter.

We begin with the first class of problems.

7.3.1 Maximization of Power with Prescribed Nulls

We will investigate the problem of maximizing the functional

$$\mathcal{J}(\psi) := \int_{\mathcal{A}} |(\mathcal{K}\psi)(\hat{\boldsymbol{x}})|^2 \, ds = \int_{S^{d-1}} \alpha(\hat{\boldsymbol{x}})^2 \, |(\mathcal{K}\psi)(\hat{\boldsymbol{x}})|^2 \, ds \tag{7.33}$$

subject to the usual input power constraint $\|\psi\|_X^2 \leq 1$. Again, α is the characteristic function of $\mathcal{A} \subset S^{d-1}$. For the additional constraints we choose β to be the characteristic function of a set $\mathcal{B} \subset S^{d-1}$ which is disjoint from the set $\mathcal{A}$ and which describes the main beam sector (or at most has an intersection with $\mathcal{A}$ of measure zero). Furthermore, we require that both sets $\mathcal{A}$ and $\mathcal{B}$ contain open subsets of S^{d-1}. Notice that if it is necessary to put a constraint on the entire region outside the main beam sector, then we may take $\beta = 1 - \alpha$. Specifically, we consider the constraint either

(a) **in integral form:**

$$g(\psi) := \int_{S^{d-1}} \beta(\hat{\boldsymbol{x}})^2 \, |(\mathcal{K}\psi)(\hat{\boldsymbol{x}})|^2 \, ds \; - \; c^2 \; \leq \; 0 \tag{7.34}$$

or

(b) **in pointwise form:**

$$g(\psi) := \sum_{j=0}^{m} w_j \, |(\mathcal{K}\psi)(\hat{\boldsymbol{x}}_j)|^2 \; - \; c^2 \; \leq \; 0 \tag{7.35}$$

where c is a prescribed (possibly small) positive constant. If we denote the constraint set, as before, by U then

$$U = \left\{ \psi \in X : \|\psi\|_X \leq 1 \text{ and } g(\psi) \leq 0 \right\}. \tag{7.36}$$

We note that this set is clearly bounded as a subset of the unit ball. It is both a convex set, since g is convex by Lemma 3.32 (a), and a closed set in X as can be checked easily using the continuity of the operator $\mathcal{K}$. Hence the general existence theorem (Theorem 3.3) concerning the maximum of the weakly sequentially continuous functional $\mathcal{J}$ over a closed and bounded convex set guarantees the existence of an optimal solution ψ^o. Many problems can be framed in this general context. In particular, we can easily see the close relationship with the classical Dolph problem in which the main beam width and power in the prescribed direction are specified and we constrain the side-lobe level outside the main beam direction. In this classical case, the side-lobe constraint can be written as

$$\max_{\hat{\boldsymbol{x}} \in \mathcal{B}} |f(\hat{\boldsymbol{x}})| \; \leq \; c \, ,$$

a pointwise constraint where c is an arbitrarily assigned constant. The present context suggests that we write the side-lobe constraint as a mean-square constraint, i.e.

$$\|f\|_{L^2(\mathcal{B})} \leq c.$$

Indeed, we will refer to the problem with this L^2-constraint as the **Generalized Dolph Problem** [13] and will analyze it in Chapter 8.

Let us start with the constraints given in integral form. The problem in this case can be treated by using the Lagrange Multiplier Rule as formulated in Theorem 3.22, where we write the norm constraint in terms of a function $h : X \to \mathbb{R}$, given by $h(\psi) := \|\psi\|_X^2 - 1$. The computations done in Section 7.1 (specifically Lemma 7.1) show that $\mathcal{J}$, g and h are Fréchet differentiable with gradients given by

$$\nabla\mathcal{J}(\psi) = 2\,\mathcal{K}^*(\alpha^2\mathcal{K}\psi)\,, \tag{7.37a}$$

$$\nabla g(\psi) = 2\,\mathcal{K}^*(\beta^2\mathcal{K}\psi)\,, \tag{7.37b}$$

$$\nabla h(\psi) = 2\,\psi\,. \tag{7.37c}$$

The Lagrange Multiplier Rule implies that there exist multipliers $\rho \geq 0$ and $\mu \geq 0$ such that

$$\nabla\mathcal{J}(\psi^o) \;-\rho\,\nabla g(\psi^o) \;-\; \mu\,\nabla h(\psi^o) = 0\,, \tag{7.38a}$$

$$\rho\, g(\psi^o) = 0\,, \tag{7.38b}$$

$$\mu\, h(\psi^o) = 0\,, \tag{7.38c}$$

provided the constraint qualifications hold i.e., for some $\psi \in X$

$$g(\psi^o) \;+\; \mathrm{Re}\left(\nabla g(\psi^o)\,,\psi\right)_X \;<\; 0\,, \tag{7.39a}$$

and

$$h(\psi^o) \;+\; \mathrm{Re}\left(\nabla h(\psi^o)\,,\psi\right)_X \;<\; 0\,. \tag{7.39b}$$

To show that the inequalities (7.39a) and (7.39b) hold for some choice of $\psi \in X$, we consider the specific forms of the functions g and h. In particular we need to show that, for some ψ,

$$\|\beta\mathcal{K}\psi^o\|^2_{L^2(S^{d-1})} \;+\; 2\,\mathrm{Re}\left(\mathcal{K}^*\beta^2\mathcal{K}\psi^o,\psi\right)_X \;<\; c^2\,, \tag{7.40a}$$

and

$$\|\psi^o\|_X^2 \;+\; 2\,\mathrm{Re}\left(\psi^o,\psi\right)_X \;<\; 1\,. \tag{7.40b}$$

To do this, we consider functions of the form $\psi = -\nu\,\psi^o$ for some $\nu > 0$. For such functions the two constraint qualifications reduce to

$$(1-2\nu)\,\|\beta\mathcal{K}\psi^o\|^2_{L^2(S^{d-1})} \;<\; c^2 \quad\text{and}\quad (1-2\nu)\,\|\psi^o\|_X^2 \;<\; 1\,,$$

which are certainly valid for every $\nu > 0$ since $\psi^o \neq 0$ and $\beta\,\mathcal{K}\psi^o \neq 0$.

It now follows from the theorem on Lagrange multipliers that there exist constants $\rho \geq 0$ and $\mu \geq 0$, for which the optimal solution, ψ^o, necessarily satisfies the equations

$$-\mathcal{K}^*(\alpha^2\mathcal{K}\psi^o) \;+\; \rho\,\mathcal{K}^*(\beta^2\,\mathcal{K}\psi^o) \;+\; \mu\,\psi^o = 0\,, \tag{7.41a}$$

$$\rho\left(\|\beta\mathcal{K}\psi^o\|^2_{L^2(S^{d-1})} - c^2\right) = 0\,, \tag{7.41b}$$

$$\mu\left(\|\psi^o\|^2_X - 1\right) = 0\,. \tag{7.41c}$$

We now examine this set of necessary conditions. First, let us consider the norm constraint (7.41c): either it is active or not. Suppose that it is not active. Then we must take $\mu = 0$ and the equation for the optimal solution becomes

$$\mathcal{K}^*(\alpha^2\mathcal{K}\psi^o) \;-\; \rho\,\mathcal{K}^*(\beta^2\mathcal{K}\psi^o) \;=\; 0$$

from which it follows that, for all $\psi \in X$,

$$\left(\mathcal{K}^*\alpha^2\mathcal{K}\psi^o - \rho\,\mathcal{K}^*\beta^2\mathcal{K}\psi^o\,,\,\psi\right)_X \;=\; 0$$

or

$$\left((\alpha^2 - \rho\,\beta^2)\mathcal{K}\psi^o, \mathcal{K}\psi\right)_X \;=\; 0 \quad \text{for all } \psi \in X\,.$$

We make now the assumption that, in addition to (A1)–(A3) there exists an open (relative to S^{d-1}) set $\mathcal{O}$ with $\mathcal{O} \subset (\operatorname{supp}\alpha) \setminus (\operatorname{supp}\beta)$. Furthermore, we assume that the range of $\mathcal{K}$ is dense in $L^2(S^{d-1})$. Then we conclude that

$$\left((\alpha^2 - \rho\,\beta^2)\mathcal{K}\psi^o, \varphi\right)_X \;=\; 0 \quad \text{for all } \varphi \in L^2(S^{d-1})\,.$$

In particular, this last equation must hold for

$$\varphi \;=\; \begin{cases} \mathcal{K}\psi^o \text{ in } \mathcal{O}\,, \\ \;0 \quad \text{in } S^{d-1} \setminus \mathcal{O}\,, \end{cases}$$

which yields

$$\int\limits_{\mathcal{O}} \alpha^2 |\mathcal{K}\psi^o|^2 ds \;=\; 0\,.$$

Assumptions (A1) and (A2) yield $\psi^o = 0$ on X as before.

We conclude that the norm constraint (7.41c) is always active and therefore that μ must be strictly positive. Now, with regard to (7.41a) let us first assume that the constraint (7.41b) is *not* active i.e., that $\|\beta\mathcal{K}\psi^o\|_{L^2(S^{d-1})} < c$. Then the second equation implies that $\rho = 0$ and the first equation becomes

$$\mathcal{K}^*(\alpha^2\mathcal{K}\psi^o) \;-\; \mu\,\psi^o \;=\; 0\,,$$

which is just the eigenvalue problem for the operator $\mathcal{K}^*\alpha^2\mathcal{K}$.

Therefore, we have restricted the class of all possible solutions of the optimization problem. Either they can be eigenfunctions ψ_j of $\mathcal{K}^*\alpha^2\mathcal{K}$, normalized by requiring $\|\psi_j\|_X = 1$, and with the additional property $\|\beta\mathcal{K}\psi_j\|_{L^2(S^{d-1})} < c$, or they can be solutions ψ of the set of equations

$$-\mathcal{K}^*(\alpha^2\mathcal{K}\psi) \;+\; \rho\,\mathcal{K}^*(\beta^2\,K\psi) \;+\; \mu\,\psi = 0\,, \tag{7.42a}$$

$$\|\beta\mathcal{K}\psi\|^2_{L^2(S^{d-1})} = c^2\,, \tag{7.42b}$$

$$\|\psi\|^2_X = 1\,. \tag{7.42c}$$

For every $\rho > 0$ we can consider (7.42a) as an eigenvalue equation for the compact operator $\mathcal{K}^*(\alpha^2\mathcal{K}\psi) - \rho\,\mathcal{K}^*(\beta^2\,K\psi)$. Therefore, we have to find $\rho > 0$ such that the normalized eigenfunctions ψ satisfy $\|\beta\,\mathcal{K}\psi\|_{L^2(S^{d-1})} = c$.

Before proceeding to a particular example, we remark that we can treat the case where $\mathcal{J}$ has the form

$$\mathcal{J}(\psi) \;=\; \sum_{j=1}^{m} w_j\,|(\mathcal{K}\psi)(\hat{\boldsymbol{x}}_j)|^2\,, \quad \psi \in X\,,$$

by the same method. The only difference is the form of the Fréchet derivative

$$\nabla\mathcal{J}(\psi) \;=\; 2\sum_{j=1}^{m} w_j\,(\mathcal{K}\psi)(\hat{\boldsymbol{x}}_j)\,p_j\,,$$

where $p_j \in X$ is the Riesz representation of the functional $\varphi \mapsto (\mathcal{K}\varphi)(\hat{\boldsymbol{x}}_j)$, $\varphi \in X$. We will not treat this case below; both the theory and computational methods can be developed following the methods already presented.

7.3.2 A Particular Example – The Line Source

As a particular example of the preceeding analysis, we consider the linear line source and will use the same notation as in Subsection 7.3.1. We choose this particular example because of its close association with finite linear arrays also discussed in [89]; indeed we show that the techniques developed here lead to the same result reported by these authors.

Specifically, using the material we developed in the preceding copy, we consider the line source and study the optimization problem in which there is a single direction in which we wish to maximize the power. At the same time, we keep the constraints in the form (7.36), so that the available surface current power is bounded (by 1) and we wish to create very low side lobes in a sector of the far field which does not include the main beam direction. We have then the optimization problem:

for fixed $\hat{\boldsymbol{x}}_0 \in S^{d-1}$,

$$\text{maximize} \quad |(\mathcal{K}\psi)(\hat{\boldsymbol{x}}_0)|^2$$

subject to

$$\|\psi\|^2_X \;\leq\; 1 \quad\text{and}\quad \|\beta\,\mathcal{K}\psi\|_{L^2(S^{d-1})} \;\leq\; c\,.$$

For this problem, as before, the constraint qualifications are satisfied and we may invoke the Lagrange Multiplier Rule. With respect to the particular case of the line source, recall that, writing the spherical coordinates of $\hat{\boldsymbol{x}}_0$ as $(\theta_0, \phi_0) \in [0, \pi] \times (0, 2\pi)$, we can introduce the operator $K : L^2(-\ell, \ell) \to L^2(-1,1)$ by

$$(K\psi)(t) \;=\; \sqrt{1-t^2} \int\limits_{-\ell}^{\ell} \psi(s)\, \mathrm{e}^{-ikst}\, ds\,, \quad |t| \le 1\,,$$

so that the optimization problem is to

$$\text{maximize} \quad \mathcal{J}(\psi) \;=\; |(K\psi)(\cos\theta_0)|^2\,, \tag{7.43a}$$

subject to

$$\|\psi\|^2_{L^2(-\ell,\ell)} \;\le\; 1 \quad \text{and} \quad \|\tilde{\beta}\, K\psi\|^2_{L^2(-1,1)} \;\le\; c^2\,, \tag{7.43b}$$

where $\tilde{\beta}(t) := \left(\int_0^{2\pi} \beta(\arccos t, \phi)^2\, d\phi\right)^{1/2}$, $|t| \le 1$.

We see that $(K\psi)(\cos\theta_0) = (\psi, p)_{L^2(-\ell,\ell)}$ where $p(s) = \sin\theta_0\, \exp(iks\cos\theta_0)$, $|s| \le \ell$. Since, by Lemma 7.1, the gradient of $\mathcal{J}$ is given by $\nabla\mathcal{J}(\psi) = 2\,(K\psi)(\cos\theta_0)\, p$ the optimal solution ψ^o of (7.43) necessarily satisfies the Lagrange equations

$$-(K\psi^o)(\cos\theta_0)\, p \;+\; \rho\, K^*(\tilde{\beta}^2 K\psi^o) \;+\; \mu\,\psi^o = 0\,, \tag{7.44a}$$

$$\rho\left(\, \|\tilde{\beta} K\psi^o\|^2_{L^2(-1,1)} - c^2\right) = 0\,, \tag{7.44b}$$

$$\mu\left(\|\psi^o\|^2_{L^2(-\ell,\ell)} - 1\right) = 0\,. \tag{7.44c}$$

We note that at least one of the constraints must be active, for if not, then necessarily both ρ and μ must vanish and (7.44a) must reduce to $p = 0$ since $(K\psi^o)(\cos\theta_0) \ne 0$. This is impossible under the hypothesis of the problem.

A phase change does not alter the optimality property of ψ^o. If we divide (7.44a) by $(K\psi^o)(\cos\theta_0)$ and define

$$\tilde{\psi}^o := \psi^o\, \mathrm{sign}\, \overline{(K\psi^o)(\cos\theta_0)} \;=\; \psi^o\, \frac{\overline{(K\psi^o)(\cos\theta_0)}}{|(K\psi^o)(\cos\theta_0)|}\,,$$

$$\tilde{\rho} \;=\; \frac{\rho}{|(K\psi^o)(\cos\theta_0)|}\,,$$

$$\tilde{\mu} \;=\; \frac{\mu}{|(K\psi^o)(\cos\theta_0)|}\,,$$

and replace $\tilde{\psi}^o$, $\tilde{\rho}$, and $\tilde{\mu}$ by ψ^o, ρ, and μ again, we arrive at the system

$$\rho\, K^*(\tilde{\beta}^2 K\psi^o) \;+\; \mu\,\psi^o = p\,, \tag{7.45a}$$

$$\rho\left(\, \|\tilde{\beta} K\psi^o\|^2_{L^2(-1,1)} - c^2\right) = 0\,, \tag{7.45b}$$

$$\mu\left(\|\psi^o\|^2_{L^2(-\ell,\ell)} - 1\right) = 0\,. \tag{7.45c}$$

We noted above that at least one of the constraints (7.45b) and (7.45c) must be active. Suppose first that the constraint (7.45b) governing the null directions is inactive. Then ρ must be zero and the multiplier equation, together with the active norm constraint, allows us to compute the solution. Indeed, in this case, ψ^o is a multiple of p, and since $\|\psi^o\|_{L^2(-\ell,\ell)} = 1$, this function must be

$$\psi^o(s) \;=\; \frac{1}{\sqrt{2\ell}}\, \mathrm{e}^{iks\cos\theta_0} \quad \text{for } s \in [-\ell, \ell]\,. \tag{7.46}$$

If, on the other hand, the norm constraint is inactive, then $\mu = 0$ and the multiplier equation becomes

$$K^*(\tilde{\beta}^2 K\psi^o) \;=\; \frac{1}{\rho}\, p \tag{7.47}$$

which is an integral equation of the first kind for ψ^o. This equation is not solvable in $L^2(-\ell, \ell)$. Indeed, we have (see (7.22))

$$(K^*\tilde{\beta}^2 K\psi)(t) \;=\; \int_{-\ell}^{\ell} b(t-s)\,\psi(s)\,ds\,, \quad |t| \le \ell\,,$$

with

$$b(\tau) \;=\; \int_{-1}^{1} (1-s^2)\,\tilde{\beta}(s)^2\, \mathrm{e}^{iks\tau}\,ds\,, \quad \tau \in \mathbb{R}\,,$$

which is the Fourier transform $\hat{h}(-k\tau)$ of

$$h(s) \;:=\; \begin{cases} (1-s^2)\,\tilde{\beta}(s)^2, & |s| \le 1\,, \\ 0\,, & |s| > 1. \end{cases}$$

Since $h \in L^1(\mathbb{R})$ we know that $B(\tau) \to 0$ as $|\tau|$ tends to infinity, which implies that $(K^*\tilde{\beta}^2 K\psi)(t) \to 0$ as $|t|$ tends to infinity, and this is a contradiction to the fact that $|p(t)|$ is constant. Therefore, the norm constraint (7.45c) must be active.

If *both* constraints are active, then (7.45a) becomes the Fredholm integral equation of the second kind

$$\big[K^*\tilde{\beta}^2 K \;+\; \gamma I\big]\psi^o \;=\; \frac{1}{\rho}\, p\,, \tag{7.48}$$

where $\gamma = \mu/\rho$. This is the result cited in [89]. The parameters γ and ρ must be determined in such a way that the solution $\psi^o = \psi^o(\gamma, \rho) \in L^2(-\ell, \ell)$ of (7.48) satisfies the equations $\|\tilde{\beta} K\psi^o\|_{L^2(-1,1)} = c$ and $\|\psi^o\|_{L^2(-\ell,\ell)} = 1$. For every $\gamma > 0$ and $\rho > 0$, the solution of (7.48) exists and is unique. Uniqueness holds as can be seen immediately by multiplying the homogeneous equation by

ψ^o thereby obtaining $\|\tilde{\beta}K\psi^o\|^2_{L^2(-1,1)} + \gamma\,\|\psi^o\|^2_{L^2(-\ell,\ell)} = 0$ and thus $\psi^o = 0$. Fredholm's alternative can now be used, much as in Section 4.5, to prove the existence of a solution.

We can reduce the problem to determine only one parameter. Indeed, we observe that the solution ψ^o of (7.48) is given by $\psi^o = \psi^{oo}/\rho$ where $\psi^{oo} = \psi^{oo}_\gamma$ solves

$$\big[K^*\tilde{\beta}^2K \;+\; \gamma I\big]\psi^{oo} \;=\; p\,. \tag{7.49a}$$

We thus have to determine ρ and γ such that

$$\|\psi^{oo}_\gamma\|_{L^2(-\ell,\ell)} \;=\; \rho \quad\text{and}\quad \|\tilde{\beta}K\psi^{oo}_\gamma\|_{L^2(-1,1)} \;=\; \rho c \;=\; c\|\psi^{oo}_\gamma\|_{L^2(-\ell,\ell)}\,.$$

The equation

$$\|\tilde{\beta}K\psi^{oo}_\gamma\|^2_{L^2(-1,1)} \;-\; c^2\|\psi^{oo}_\gamma\|^2_{L^2(-\ell,\ell)} \;=\; 0 \tag{7.49b}$$

is one equation for the unknown parameter γ.

It should be pointed out here that we *cannot* make the far field pattern vanish on all of the set $\mathcal{B}$ because of the analyticity of the far field pattern. Indeed, the assumption that the set $\mathcal{B}$ contains an open set would force the far field pattern to vanish everywhere if it vanished on all of $\mathcal{B}$. This observation raises the question of whether we can force nulls at a discrete set of points. This question is addressed by considering the problem with pointwise constraints which we will do in the next subsection.

Example 7.8. We finish this part by presenting a numerical example for which β is the characteristic function of the strip $\big([0,\pi/3]\cup[2\pi/3,\pi]\big)\times[0,2\pi]$ i.e. $\tilde{\beta}$ has the form

$$\tilde{\beta}(\cos\theta) \;=\; \begin{cases} \sqrt{2\pi}\,, & \theta\in[0,\pi/3]\cup[2\pi/3,\pi]\,,\\ 0\,, & \text{otherwise}\,.\end{cases}$$

Furthermore, we take $\ell = 1$ and $\cos\theta_0 = 0$. In this case, the Riesz representation is $p \equiv 1$. We solved (7.49b) for γ by the Regula falsi. For the wave length $\lambda = 2$ and levels $c = 0.1$ and $c = 0.5$ we show in Figure 7.5 the plots of $|f| = |K\psi^o|$ in comparison to the pattern $|f|$ corresponding to the uniform feeding $\psi \equiv 1/\sqrt{2}$.

7.3.3 Pointwise Constraints

We go back to the general case where we wish to maximize the power in a given sector, subject to the usual norm constraint, but now with the constraint (see (7.35) above),

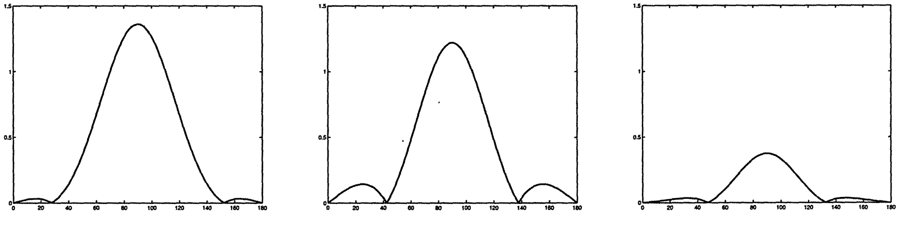


Fig. 7.5. $\lambda = 2$: $\theta \mapsto |f(\theta)|$ for $\psi \equiv 1/\sqrt{2}$ (left) and for ψ^o corresponding to $c = 0.5$ (center) and $c = 0.1$ (right)

$$g(\psi) \;:=\; \sum_{j=0}^{m} w_j \, |(\mathcal{K}\psi)(\hat{\boldsymbol{x}}_j)|^2 \;-\; c^2 \;\leq\; 0\,, \tag{7.50}$$

The constraint qualifications are checked as before, and we will not do the details here. Recall, moreover, that we have computed the Fréchet differential of the functional g in Section 7.2 and have seen that the gradient is

$$\nabla g(\psi) \;=\; 2\sum_{j=1}^{m} w_j \left[(\mathcal{K}\psi)(\hat{\boldsymbol{x}}_j)\right] p_j\,, \tag{7.51}$$

where the p_j are given by the Riesz Representation Theorem,

$$(\mathcal{K}\varphi)(\hat{\boldsymbol{x}}_j) \;=\; (\varphi, p_j)_X\,, \quad \varphi \in X\,.$$

The multiplier equations then have the form

$$-\mathcal{K}^*(\alpha^2\mathcal{K}\psi^o) \;+\; \rho\sum_{j=1}^{m} w_j (\mathcal{K}\psi^o)(\hat{\boldsymbol{x}}_j)\, p_j \;+\; \mu\,\psi^o = 0\,, \tag{7.52a}$$

$$\rho\left(\sum_{j=1}^{m} w_j |(\mathcal{K}\psi^o)(\hat{\boldsymbol{x}}_j)|^2 \;-\; c^2\right) = 0\,, \tag{7.52b}$$

$$\mu\left(\|\psi^o\|_X^2 \;-\; 1\right) = 0\,. \tag{7.52c}$$

As in the preceeding section, it is necessary to consider the three cases where one or more of the constraints is active. But now different arguments are needed to handle the new situation. In the case that the first constraint is inactive, ρ has to vanish and the problem reduces to the now familiar eigenvalue problem for the operator $\mathcal{K}^*\alpha^2\mathcal{K}$. If the first is active, but the norm constraint is not, then $\mu = 0$ and the problem becomes

$$\mathcal{K}^*(\alpha^2\mathcal{K}\psi^o) \;=\; \rho\sum_{j=1}^{m} w_j (\mathcal{K}\psi^o)(\hat{\boldsymbol{x}}_j)\, p_j\,.$$

It can be shown that this case cannot occur if the range of $\mathcal{K}$ is dense in $C(S^{d-1})$. Indeed, multiplication of this equation by some $\varphi \in X$ yields

$$\left(\alpha\mathcal{K}\psi^o, \alpha\mathcal{K}\varphi\right)_{L^2(S^{d-1})} \;=\; \rho\sum_{j=1}^{m} w_j (\mathcal{K}\psi^o)(\hat{\boldsymbol{x}}_j)\,\overline{(\mathcal{K}\varphi)(\hat{\boldsymbol{x}}_j)}\,.$$

Now, assuming $\varphi^o \neq 0$, we choose $\varphi \in X$ such that $(\mathcal{K}\varphi)(\hat{\boldsymbol{x}}_j)$ is small for all j but $\mathcal{K}\varphi$ is close to $\mathcal{K}\psi^o$ with respect to the L^2–norm. A careful analysis leads to a contradiction (see proof of Theorem 7.12 below). Thus $\varphi^o = 0$ which clearly is not optimal.

The final case is the one in which both constraints are active so that both multipliers in (7.52a)–(7.52c) must be found: For each $\rho > 0$ we determine the eigenvalues $\mu_j = \mu_j(\rho)$ and normalized eigenfunctions $\psi_j = \psi_j(\rho)$ of the operator

$$\mathcal{K}^*\alpha^2\mathcal{K} \;-\; \rho\sum_{j=1}^{m} w_j (\mathcal{K}\cdot)(\hat{\boldsymbol{x}}_j)\, p_j$$

and there determine ρ such that $g(\psi_j) = 0$.

7.3.4 Minimization of Pattern Perturbation

In the preceeding three subsections we have considered the maximization of power in one direction while minimizing it in other prescribed regions of the far field. We now turn to the second problem. For this problem we wish to modify an existing (and ostensibly desirable) far field pattern to reduce side lobes at prescribed locations or in specified sectors of the far field, while preserving as closely as possible the existing pattern outside these given positions. In this way, one attempts to preserve (to the extent possible) desirable pattern characteristics as, for example, gain and beam width. The problem has been addressed by [123], [128] and [79] (see also [52]).

In our context, the problem is to minimize the functional

$$\mathcal{J}(\psi) \;:=\; \left\|\mathcal{K}\psi - f^o\right\|^2_{L^2(\mathcal{A})} \;=\; \left\|\alpha[\mathcal{K}\psi - f^o]\right\|^2_{L^2(S^{d-1})} \tag{7.53}$$

where the function α is the characteristic function of the part $\mathcal{A} \subset S^{d-1}$ of the far field which is to be preserved. Then the sector $\mathcal{B} \subset S^{d-1}$ where we wish to minimize the pattern is described by the characteristic function β which can, e.g., be $1 - \alpha$. As usual, we impose a power bound on the inputs to the antenna $\|\psi\|_X \leq 1$.

As in the previous case, we take the constraints either

(a) **in integral form**:

$$g(\psi) \;:=\; \int_{S^{d-1}} \beta(\hat{\boldsymbol{x}})^2\,|(\mathcal{K}\psi)(\hat{\boldsymbol{x}})|^2\,ds(\hat{\boldsymbol{x}}) \;-\; c^2 \;\leq\; 0\,, \tag{7.54}$$

or

(b) **in pointwise form:**

$$g(\psi) \; := \; \sum_{j=1}^{m} w_j \, |(\mathcal{K}\psi)(\hat{\boldsymbol{x}}_j)|^2 \; - \; c^2 \; \leq \; 0 \, , \tag{7.55}$$

where again $c > 0$ is a preassigned constant.

Rather than discussing this problem first for the integral constraints (7.54) and then for the pointwise case (7.55) as we did in the previous section, we will present an analysis of a problem containing the two types of constraints.

Let us denote the given far field by f^o. We are interested in the case for which f^o has been established and is the pattern we wish to maintain as closely as possible in the mean-square sense. Thus, the function f^o is in the range of the compact operator $\mathcal{K}$, i.e., $\mathcal{K}\varphi^o = f^o$ for some $\varphi^o \in X$, and consequently the trivial estimate

$$\|\mathcal{K}\psi - f^o\|_{L^2(S^{d-1})} \; = \; \|\mathcal{K}(\psi - \varphi^o)\|_{L^2(S^{d-1})} \; \leq \; \|\mathcal{K}\| \; \|\psi - \varphi^o\|_X$$

shows that by making $\|\psi - \varphi^o\|_X$ small, we make only small perturbations in the far field[3]. Moreover, let us suppose, as if often the case, that we insist on maintaining the same level of power in the direction θ_0 of the main beam, that is we want to impose the particular constraint

$$|(\mathcal{K}\psi)(\hat{\boldsymbol{x}}_0)| \; = \; |(\mathcal{K}\varphi^o)(\hat{\boldsymbol{x}}_0)| \, .$$

If we denote the sector (connected or not) of the far field where we wish to minimize the side lobe level by $\mathcal{B}$, and its characteristic function by β as before, then we may take

$$\mathcal{J}(\psi) \; := \; \mu \, \|\psi - \varphi^o\|_X^2 \; + \; \|\beta\mathcal{K}\psi\|_{L^2(S^{d-1})}^2 \, , \tag{7.56}$$

as the cost functional and pose the optimization problem:

$$\text{Minimize} \quad \mathcal{J}(\psi) \quad \text{subject to} \quad |(\mathcal{K}\psi)(\hat{\boldsymbol{x}}_0)| \; = \; c \, . \tag{7.57}$$

Here the constant c is just defined by the absolute value of the given far field i.e., $c = |f^o(\hat{\boldsymbol{x}}_0)| = |(\mathcal{K}\varphi^o)(\hat{\boldsymbol{x}}_0)|$ and $\mu > 0$ is a coupling parameter. The corresponding problem for finite arrays was discussed e.g., by Shore [123] as well as in [128]. The problem is related to problems of adaptive antenna arrays which are discussed recently in [109]. Most of these applications are subsumed here.

[3] Recall that the quality factor of an antenna, as defined earlier (see (3.40)) is related to the norm $\|\mathcal{K}\|$, by $\inf_{\varphi \in X} Q(\varphi) = \|\mathcal{K}\|^{-2}$, and so the gain of the antenna is bounded by $\|\mathcal{K}\|^2$. In so far as the far field operator intrinsically models the physical character of the radiating structure, we see clearly that the gain is limited by the physical nature of the radiating structure.

One should, of course, start with the problem of the existence of an optimal solution. We note that the functional $\mathcal{J}$ is uniformly convex. Indeed, the binomial formula yields immediately uniform convexity of the first term while application of Lemma 7.1 yields convexity of the second term. On the other hand, the admissible set U, given by

$$U \;=\; \bigl\{\psi \in X : |(\mathcal{K}\psi)(\hat{\boldsymbol{x}}_0)| = c\bigr\}\,,$$

fails to be either convex or bounded. However, this set is still weakly sequentially closed due to the compactness of $\mathcal{K} : X \to C(S^{d-1})$. Therefore, the general existence Theorem 3.3 is applicable and yields the existence of optimal solutions $\psi^o \in U$ of (7.57).

Knowing that solutions exist, we now discuss the use of necessary conditions for this constrained minimization problem and apply the Lagrange multiplier rule to (7.57). To find the correct form for the multiplier rule, we must, as usual, compute the Fréchet derivatives of the functional $\mathcal{J}$ and of the constraint function $h(\psi) := |(\mathcal{K}\psi)(\hat{\boldsymbol{x}}_0)|^2 - c^2$. We have done this in the previous subsections, and using those results it follows that there exists a multiplier $\rho \in \mathbb{R}$ for which the optimal solution, ψ^o must satisfy

$$\mu\,(\psi^o - \varphi^o) \;+\; \mathcal{K}^*(\beta^2\mathcal{K}\psi^o) \;+\; \rho\,(\mathcal{K}\psi^o)(\hat{\boldsymbol{x}}_0)\,p \;=\; 0 \qquad (7.58)$$

where p is given by the Riesz representation of the map $\varphi \mapsto (\mathcal{K}\varphi)(\hat{\boldsymbol{x}}_0)$.

This last equation (7.58) we can rewrite as an operator equation of Fredholm type, namely

$$\bigl[\mu I \;+\; \mathcal{K}^*\beta^2\mathcal{K}\bigr]\psi^o \;=\; \mu\varphi^o \;-\; \nu p\,, \qquad (7.59)$$

where the (complex) parameter ν is just

$$\nu \;:=\; \rho\,(\mathcal{K}\psi^o)(\hat{\boldsymbol{x}}_0)\,. \qquad (7.60)$$

Linearity of the equation (7.59) allows us to split the problem into simpler parts

$$\bigl[\mu I \;+\; \mathcal{K}^*\beta^2\mathcal{K}\bigr]\psi = \mu\,\varphi^o \quad \text{and} \qquad (7.61\text{a})$$

$$\bigl[\mu I \;+\; \mathcal{K}^*\beta^2\mathcal{K}\bigr]\psi = p \qquad (7.61\text{b})$$

whose solutions we denote, respectively, by ψ_1 and ψ_2. Then the solution of (7.59) is given by

$$\psi^o \;=\; \psi_1 \;-\; \nu\,\psi_2\,.$$

Our problem then reduces to finding appropriate values of the parameter ν which is related to the unknown Lagrange multiplier $\rho \in \mathbb{R}$ by (7.60). It is here that we use the constraint equation $|(\mathcal{K}\psi^o)(\hat{\boldsymbol{x}}_0)| = c$. To this end, let us define z_1 and z_2 by

$$z_{1'} \;:=\; (\mathcal{K}\psi_1)(\hat{\boldsymbol{x}}_0) \quad \text{and} \quad z_2 \;:=\; (\mathcal{K}\psi_2)(\hat{\boldsymbol{x}}_0)\,. \qquad (7.62)$$

Then the constraint equation $|(\mathcal{K}\psi^o)(\hat{\boldsymbol{x}}_0)| = c$ can obviously be rewritten as $\big|z_1 - \nu\, z_2\big| = c$. Recalling the definition (7.60) of ν, it follows that ν satisfies the relation $\nu = \rho\,(z_1 - \nu\, z_2)$ which can be solved for ν yielding

$$\nu \;=\; \frac{\rho\, z_1}{1\,+\,\rho\, z_2}\,.$$

We now use the *definition* of ψ_2 as a solution of (7.61b) to find that we can write

$$\begin{aligned} z_2 = (\mathcal{K}\psi_2)(\hat{\boldsymbol{x}}_0) \;&=\; (\psi_2, p)_X \;=\; \big(\psi_2\,,\,(\mu I + \mathcal{K}^*\beta^2\mathcal{K})\psi_2\big)_X \\ = \mu\,\|\psi_2\|_X^2 \;&+\; \|\beta\,\mathcal{K}\psi_2\|^2_{L^2(S^{d-1})}\,. \end{aligned} \tag{7.63}$$

We see that z_2 is real-valued and strictly positive. Substituting the form of ν, the constraint relation $\big|z_1 - \nu\, z_2\big| = c$ can then be rewritten as $|1 + \rho\, z_2| = |z_1|/c$, i.e.,

$$\rho \;=\; \frac{1}{z_2}\left[\frac{|z_1|}{c} \;-\; 1\right] \quad \text{or} \quad \rho \;=\; -\frac{1}{z_2}\left[\frac{|z_1|}{c} \;+\; 1\right]\,.$$

This results in

$$\nu_\pm \;=\; \frac{z_1}{z_2}\left(1 \;\pm\; \frac{c}{|z_1|}\right). \tag{7.64}$$

Since z_1 and z_2 are known from the solutions of (7.61a), (7.61b) we find that there are *two* solutions of the multiplier rule which are candidates for the optimal solution of the original problem, namely

$$\psi^o_\pm \;=\; \psi_1 \;-\; \frac{z_1}{z_2}\left(1 \;\pm\; \frac{c}{|z_1|}\right)\psi_2\,. \tag{7.65}$$

We remark that this general analysis does *not* depend on any specific form of the antenna. However, to proceed to numerical results, we need to find specific forms for the operator $\mathcal{K}$ and the function p whose existence is given by the Riesz theorem.

Example 7.9. Returning to Example 7.6, consider a circular line source of radius $a > 0$. Let φ^o be the solution determined in Example 7.6, i.e. the solution of the problem of maximizing the power $\int_{\alpha_1}^{\alpha_2} |K\varphi|^2 ds$ subject to $\|\varphi\|_{L^2(0,2\pi)} \leq 1$.

We choose β to be the characteristic function of some interval $[\beta_1, \beta_2] \subset [0, 2\pi]$ which is disjoint of (α_1, α_2). The operator $K^*\beta^2 K$ has then the same form as the operator $K^*\alpha^2 K$, i.e.

$$(K^*\beta^2 K\psi)(t) \;=\; \sum_{\ell,m\in\mathbb{Z}} b_{\ell m}\,\psi_m\,\mathrm{e}^{i\ell t}\,, \quad 0 \leq t \leq 2\pi\,,$$

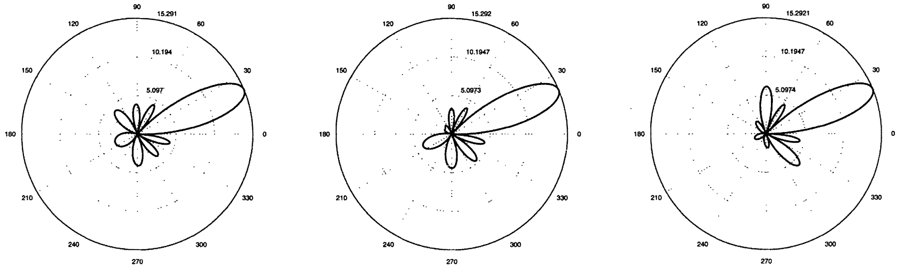
where

$$\psi_m = \frac{1}{2\pi} \int\limits_0^{2\pi} \psi(t)\, \mathrm{e}^{-imt}\, dt\,, \quad m \in \mathbb{Z}\,,$$

are the Fourier coefficients of $\psi \in L^2(0, 2\pi)$ and

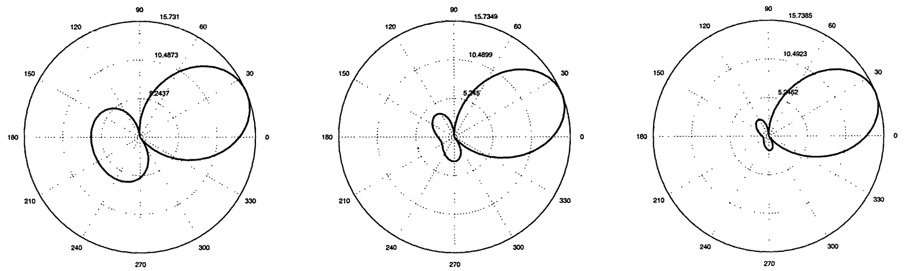
$$b_{\ell m} = \begin{cases} 2\pi\, i^{m-\ell}\, J_m(ka)\, J_\ell(ka)\, \dfrac{\mathrm{e}^{i(m-\ell)\beta_2} - \mathrm{e}^{i(m-\ell)\beta_1}}{i\,(m-\ell)}\,, & \ell \neq m\,, \\ 2\pi\, J_m(ka)^2\, (\beta_2 - \beta_1)\,, & \ell = m\,. \end{cases}$$

As a particular numerical experiment we take $[\alpha_1, \alpha_2] = [0, \pi/4]$, $\theta_0 = (\alpha_1 + \alpha_2)/2 = \pi/8$, and $\lambda = 1$ (Figure 7.6) or $\lambda = \pi$ (Figure 7.7), respectively, as in Example 7.6. The left plot shows the far field pattern without placing nulls (compare with Figure 7.1). The table below the plots show the parameter values for each run. For comparison, in the first column we have taken the maximal value $\max\bigl\{\|\alpha K\varphi\|^2_{L^2(0,2\pi)} : \|\varphi\|_{L^2(0,2\pi)} \leq 1\bigr\}$ without placing nulls as the reference pattern (see also Figure 7.1).



null sector:		[120°, 150°]	[120°, 300°]
μ		1	1
radiated power in [0°, 45°]	2.3949	2.3673	2.2344

Fig. 7.6. Parameters for Null Placement Problem, $\lambda = 1$



null sector:		$[120^o, 300^o]$	$[120^o, 300^o]$
μ		3	1
radiated power in $[0^o, 45^o]$	4.4750	4.3988	4.3813

Fig. 7.7. Parameters for Null Placement Problem, $\lambda = \pi$

7.4 The Optimization of Signal-to-Noise Ratio and Directivity

We now turn to a consideration of the problem of optimizing the signal-to-noise ratio subject to a constraint on the Q-factor (see Section 3.4):
For fixed $\hat{\boldsymbol{x}}_0 \in S^{d-1}$,

$$\text{maximize} \qquad \mathit{SNR}(\hat{\boldsymbol{x}}_0) \;:=\; \frac{|(\mathcal{K}\psi)(\hat{\boldsymbol{x}}_0)|^2}{\int_{S^{d-1}} \omega(\hat{\boldsymbol{y}})^2 |\mathcal{K}\psi(\hat{\boldsymbol{y}})|^2\, ds} \tag{$\mathcal{P}$}$$

over the set

$$U \;:=\; \left\{ \psi \in X \setminus \{0\} \;:\; Q(\psi) := \frac{\|\psi\|_X^2}{\|\mathcal{K}\psi\|^2_{L^2(S^{d-1})}} \leq c \right\}. \tag{7.66}$$

In the definition of the signal-to-noise ratio, ω represents the noise distribution whose support is assumed to contain a set open relative to S^{d-1}. Furthermore, we assume that the operator $\mathcal{K} : X \to C(S^{d-1})$ defined on some Hilbert space X satisfies the conditions (A1)–(A2) from the beginning of this chapter. Again, we think of X being a space of functions defined on the structure Γ.

The denominator of the signal-to-noise ratio $\mathit{SNR}(\hat{\boldsymbol{x}}_0)$ vanishes only if the function $\omega\,\mathcal{K}\psi$ vanishes almost everywhere on S^{d-1}. Our assumptions (A1), (A2) on the operator $\mathcal{K}$, specifically those which guarantee that $\mathcal{K}\psi$ is analytic, prevent this from occurring for any $\psi \neq 0$. We write the numerator in the form $(\mathcal{K}\psi)(\hat{\boldsymbol{x}}_0) = (\psi, p)_X$, $\psi \in X$, where $p \in X$ denotes the Riesz representation of $\psi \mapsto (\mathcal{K}\psi)(\hat{\boldsymbol{x}}_0)$.

As before, we will find it convenient to denote the set of functions satisfying the constraint (7.66) by U. We note that, in contrast to most of our optimization problems discussed so far, this set in neither convex nor bounded. Therefore, our general results on existence are not directly applicable.

For each fixed value $\hat{x}_0 \in S^{d-1}$, the signal-to-noise ratio, *SNR*, defines a functional of ψ which we denote by $SNR(\psi)$. Hopefully this abuse of notation will cause no confusion.

7.4.1 The Existence of Optimal Solutions

The basic existence result for the maximization of the functional

$$SNR(\psi) \;:=\; \frac{|(\mathcal{K}\psi)(\hat{x}_0)|^2}{\|\omega\,\mathcal{K}\psi\|^2_{L^2(S^{d-1})}} \;=\; \frac{|(p,\psi)_X|^2}{\|\omega\,\mathcal{K}\psi\|^2_{L^2(S^{d-1})}} \tag{7.67a}$$

on the set

$$U \;:=\; \{\psi \in X : Q(\psi) \;\leq\; c\} \tag{7.67b}$$

where Q is defined by (7.66) can be stated succinctly as:

Theorem 7.10. *If there is any non-trivial $\psi \in X$ satisfying the constraint*

$$Q(\psi) \;:=\; \frac{\|\psi\|^2_X}{\|\mathcal{K}\psi\|^2_{L^2(S^{d-1})}} \;\leq\; c, \tag{7.68}$$

then $v^o := \sup\{SNR(\psi) \,:\, \psi \neq 0\,,\; Q(\psi) \leq c\}$ is finite and there exists some admissible $\psi^o \in X$ such that $SNR(\psi^o) = v^o$ i.e., problem $(\mathcal{P})$ is solvable.

Proof: The operator $\mathcal{K} \,:\, X \to C(S^{d-1}) \subset L^2(S^{d-1})$ is compact. By the definition of v^o, there exists a maximizing sequence,

$$\{\psi_n\}_{n=1}^{\infty} \subset X,$$

such that $Q(\psi_n) \leq c$, for all $n = 1, 2, \ldots$, and for which

$$SNR(\psi_n) \longrightarrow v^o \text{ as } n \to \infty\,.$$

Note that we have not excluded the possibility that $v^o = \infty$.

Our next observation is that both *SNR* and Q are homogeneous of degree zero i.e., $SNR(z\psi) = SNR(\psi)$, and $Q(z\psi) = Q(\psi)$ for any $z \in \mathbb{C}$ and $\psi \neq 0$. With this observation, we can replace the functions ψ_n by the functions $\hat{\psi}_n := \psi_n/\,\|\psi_n\|_X$ of unit norm. Then

$$SNR(\hat{\psi}_n) \;\longrightarrow\; v^o \quad\text{and}\quad Q(\hat{\psi}_n) \;\leq\; c,\; \|\hat{\psi}_n\|_X \;=\; 1\,.$$

Since the unit ball is weakly compact in a Hilbert space, we can assume that there exists a subsequence of $\{\hat{\psi}_n\}$ which we again denote by $\{\hat{\psi}_n\}$ which converges weakly in X to some function $\psi^o \in X$ with norm $\|\psi^o\|_X \leq 1$. It follows from the compactness of the operator $\mathcal{K}$ having range in $C(S^{d-1})$, that

$$\mathcal{K}\hat{\psi}_n \longrightarrow \mathcal{K}\psi^o \quad \text{uniformly on } S^{d-1} \text{ as } n \to \infty. \tag{7.69}$$

In fact, along this normalized maximizing sequence we have

$$Q(\hat{\psi}_n) = \frac{1}{\|\mathcal{K}\hat{\psi}_n\|^2_{L^2(S^{d-1})}} \leq c,$$

which implies that

$$\|\mathcal{K}\psi^o\|^2_{L^2(S^{d-1})} \geq \frac{1}{c} \tag{7.70}$$

since also $\mathcal{K}\hat{\psi}_n \to \mathcal{K}\psi^o$ in $L^2(S^{d-1})$. Thus, in particular, $\psi^o \neq 0$ and, moreover,

$$Q(\psi^o) = \frac{\|\psi^o\|^2_X}{\|\mathcal{K}\psi^o\|^2_{L^2(S^{d-1})}} \leq \frac{1}{\|\mathcal{K}\psi^o\|^2_{L^2(S^{d-1})}} \leq c,$$

so that $\psi^o \in U$.

Now the uniform convergence in (7.69) implies convergence of the functions $\omega\,\mathcal{K}\hat{\psi}_n$ in $L^2(S^{d-1})$ and in particular, $\|\omega\,\mathcal{K}\hat{\psi}_n\|_{L^2(S^{d-1})} \to \|\omega\,\mathcal{K}\psi^o\|_{L^2(S^{d-1})}$. Therefore, $\|\omega\,\mathcal{K}\psi^o\|_{L^2(S^{d-1})} > 0$, since otherwise, the analytic function $\mathcal{K}\psi^o$ would necessarily vanish on the support of ω and therefore, by analytic continuation, everywhere on S^{d-1} by virtue of our assumption that the support contains an open set. This would contradict the fact that $\psi^o \neq 0$ (see (7.70)). Finally, by the weak sequential continuity of the functional *SNR* (see 3.30),

$$\mathit{SNR}(\hat{\psi}_n) = \frac{\left|(p, \hat{\psi}_n)_X\right|^2}{\|\omega\,\mathcal{K}\hat{\psi}_n\|^2_{L^2(S^{d-1})}} \longrightarrow \frac{\left|(p, \psi^o)_X\right|^2}{\|\omega\,\mathcal{K}\psi^o\|^2_{L^2(S^{d-1})}} = \mathit{SNR}(\psi^o).$$

In particular, $v^o = \mathit{SNR}(\psi^o) < \infty$, and the proof is complete. □

Remark: We should say something concerning the hypothesis that the set of admissible inputs, U, is non-empty. Recall, as remarked in Chapter 3, equation (3.42), that $\inf_{\psi \in X} Q(\psi) = 1/\,\|\mathcal{K}\|^2$. Hence, for every $c > 1/\,\|\mathcal{K}\|^2$ there exists a function $\hat{\psi} \in X$ with norm 1 such that $Q(\hat{\psi}) < c$. For these values of c the set U is non-empty.

7.4.2 Necessary Conditions

We are now interested in developing necessary conditions for this problem in the form of a Lagrange Multiplier rule just as we have done with the other problems discussed in the preceeding sections of this chapter. To do so, it is necessary to make a further assumption. For its formulation we recall that $p \in X$ denotes the Riesz representation of the functional $\psi \mapsto (\mathcal{K}\psi)(\hat{\boldsymbol{x}}_0)$. Then we require that

(A5) $\mathcal{K}^*(\omega^2\mathcal{K}\psi^o) \notin \operatorname{span}\{p\}$, i.e. $\mathcal{K}^*(\omega^2\mathcal{K}\psi^o)$ is not a multiple of p.

We remark that we have discussed similar situations in Chapter 1. In particular, we can give a condition which will guarantee that (A5) is satisfied.

Lemma 7.11. *If the range of the operator $\mathcal{K} : X \to C(S^{d-1})$ is dense in $C(S^{d-1})$ then (A5) is satisfied.*

Proof: Assume, on the contrary, that $\mathcal{K}^*(\omega^2\mathcal{K}\psi^o) = \mu p$ for some $\mu \in \mathbb{C}$. We note that $\mu \neq 0$ since $\|\omega\mathcal{K}\psi^o\|_{L^2(S^{d-1})} > 0$. Multiplication with any $\varphi \in X$ yields

$$\left(\omega\mathcal{K}\psi^o, \omega\mathcal{K}\varphi\right)_{L^2(S^{d-1})} \;=\; \left(\mathcal{K}^*(\omega^2\mathcal{K}\psi^o), \varphi\right)_X \;=\; \mu\,(p,\varphi)_X \;=\; \mu\,\overline{(\mathcal{K}\varphi)(\hat{\boldsymbol{x}}_0)}\,.$$

Let γ be a continuous function on S^{d-1} such that $\left|\mu\,\gamma(\hat{\boldsymbol{x}}_0)\right| = 1$ and $\left|\left(\omega\mathcal{K}\psi^o, \omega\gamma\right)_{L^2(S^{d-1})}\right| < 1/3$. Since the range of $\mathcal{K}$ is dense we can find $\varphi \in X$ with

$$\|\mathcal{K}\varphi - \gamma\|_{C(S^{d-1})} \;\le\; \frac{1}{3}\,\min\left\{\frac{1}{|\mu|}\,,\;\left[\sqrt{\int_{S^{d-1}} \omega^2\,ds}\;\left\|\omega^2\mathcal{K}\psi^o\right\|_{L^2(S^{d-1})}\right]^{-1}\right\}.$$

Then

$$\begin{aligned}
1 = \left|\mu\,\gamma(\hat{\boldsymbol{x}}_0)\right| \;&\le\; |\mu|\left|\gamma(\hat{\boldsymbol{x}}_0) - (\mathcal{K}\varphi)(\hat{\boldsymbol{x}}_0)\right| \;+\; |\mu|\left|(\mathcal{K}\varphi)(\hat{\boldsymbol{x}}_0)\right| \\
&\le |\mu|\,\|\mathcal{K}\varphi - \gamma\|_{C(S^{d-1})} \;+\; \left|\left(\omega\mathcal{K}\psi^o, \omega\mathcal{K}\varphi\right)_{L^2(S^{d-1})}\right| \\
&\le \frac{1}{3} \;+\; \left|\left(\omega\mathcal{K}\psi^o, \omega\gamma\right)_{L^2(S^{d-1})}\right| \;+\; \|\omega\mathcal{K}\psi^o\|_{L^2(S^{d-1})}\|\omega(\mathcal{K}\varphi - \gamma)\|_{L^2(S^{d-1})} \\
&< 1\,,
\end{aligned}$$

a contradiction! This ends the proof. □

The preceeding lemma is useful as there are a number of situations in which we can verify that the range of K is indeed dense, see for example [30, Theorem 5.17]. We begin the discussion of the necessary conditions by rewriting the constraint (7.68) in the form

$$g(\psi) \;:=\; \|\psi\|_X^2 - c\,\|\mathcal{K}\psi\|^2_{L^2(S^{d-1})} \;\le\; 0\,. \tag{7.71}$$

The first observation that we make is that the constraint is active on any optimal solution.

Theorem 7.12. *Let $\psi^o \in U$ be an optimal solution of the problem $(\mathcal{P})$. Then, under the assumption (A5) necessarily, $g(\psi^o) = 0$, i.e. $Q(\psi^o) = c$.*

Proof: Assume that $Q(\psi^o) < c$. We use the Fréchet derivative of the *SNR*–functional. Recalling the results of Lemma 3.32 the usual computation shows that the Frèchet derivative of *SNR* is given by

$$\begin{aligned} \mathit{SNR}'(\psi^o)\varphi &= \frac{2\,\mathrm{Re}\left[\overline{(\psi^o,p)_X}\,(\varphi,p)_X\right]}{\|\omega\,\mathcal{K}\psi^o\|^2_{L^2(S^{d-1})}} \;-\; v^o\,\frac{2\,\mathrm{Re}\,(\omega\,\mathcal{K}\varphi,\omega\,\mathcal{K}\psi^o)_{L^2(S^{d-1})}}{\|\omega\,\mathcal{K}\psi^o\|^2_{L^2(S^{d-1})}} \\ &= \frac{2}{\|\omega\,\mathcal{K}\psi^o\|^2_{L^2(S^{d-1})}}\,\mathrm{Re}\left(\varphi\,,\,(\psi^o,p)_X\,p - v^o\,\mathcal{K}^*\omega^2\mathcal{K}\psi^o\right) . \end{aligned}$$

where $v^o = \mathit{SNR}(\psi^o)$. By assumption (A5) and the facts that $(\psi^o,p)_X = (\mathcal{K}\psi^o)(\hat{x}_0) \neq 0$ and $v^o \neq 0$ we conclude that $(\psi^o,p)_X\,p - v^o\,\mathcal{K}\omega^2\mathcal{K}\psi^o \neq 0$. Therefore, there exists $\varphi \in X$ with the property that $\mathit{SNR}'(\psi^o)\varphi > 0$ and thus $\mathit{SNR}(\psi^o + \epsilon\varphi) > \mathit{SNR}(\psi^o)$ for sufficiently small $\epsilon > 0$. Moreover, since the constraint functional Q is continuous, $Q(\psi^o + \epsilon\varphi) \leq c$ for sufficiently small $\epsilon > 0$, so that $\psi^o + \epsilon\varphi$ is admissible. We may conclude that the function ψ^o is not optimal, which is a contradiction, and the proof is complete. □

As before, we will use the Lagrange multiplier rule (Theorem 3.22) to calculate solutions of the problem and, to do so, we will require that a constraint qualification is satisfied. It is necessary to impose the further, but quite mild condition, that

(A6) The constant $1/c$ is not an eigenvalue of the operator $\mathcal{K}^*\mathcal{K}$.

Remark: As the operator $\mathcal{K}^*\mathcal{K}$ is compact and therefore has a discrete spectrum, it is possible to arrange the constraint so that (A6) is satisfied provided we have some specific information on the operator $\mathcal{K}$ whose spectrum depends, of course, on the particular choice of X.

As the constraint function g, given in (7.71), is

$$g(\psi) \;=\; \|\psi\|^2_X - c\,\|\mathcal{K}\psi\|^2_{L^2(S^{d-1})}\,, \quad \psi \in X\,,$$

we need to check that the generalized Slater condition:

$$\text{there exists a } \varphi \in X \text{ such that} \quad g(\psi^o) \;+\; g'(\psi^o)\varphi \;<\; 0 \tag{7.72}$$

is satisfied where again g' is the Fréchet derivative of g. However, from Lemma 7.1 and the fact that the constraint is active at an optimal solution, we may rewrite the left-hand side of the equality (7.72) in the form

$$\begin{aligned} g(\psi^o) \;+\; g'(\psi^o)\varphi &= 2\,\mathrm{Re}\,(\varphi,\psi^o)_X \;-\; 2c\,\mathrm{Re}\,(\mathcal{K}\varphi\mathcal{K}\psi^o)_{L^2(S^{d-1})} \\ &= 2c\,\mathrm{Re}\left(\varphi\,,\,c^{-1}\psi^o - \mathcal{K}^*\mathcal{K}\psi^o\right)_X \end{aligned}$$

which does not vanish identically by assumption (A6). Hence for some $\varphi \neq 0$ we see that (7.72) is satisfied.

We have already computed the Fréchet derivative of the SNR functional, and so we have the gradients of both the objective and constraint functionals. With these gradients in hand we may write the Lagrange necessary conditions for this constrained problem as follows, where $\lambda \geq 0$ and $a := \|\omega\mathcal{K}\psi^o\|^2_{L^2(S^{d-1})}$:

$$-\frac{1}{a}\left[(\psi^o,p)_X\,p \;-\; v^o\,\mathcal{K}^*\omega^2\mathcal{K}\psi^o\right] \;+\; \lambda\left[\psi^o \;-\; c\mathcal{K}^*\mathcal{K}\psi^o\right] \;=\; 0\,, \qquad (7.73)$$

i.e.

$$\lambda\big[I \;-\; c\mathcal{K}^*\mathcal{K}\big]\psi^o \;+\; \frac{v^o}{a}\mathcal{K}^*\omega^2\mathcal{K}\psi^o \;=\; \frac{1}{a}\,(\psi^o,p)_X\,p\,. \qquad (7.74)$$

We first observe that it is not possible for the multiplier λ to vanish. Indeed, were this to be the case, then the optimal solution ψ^o would have to satisfy the equation

$$v^o\,\mathcal{K}^*\omega^2\mathcal{K}\psi^o \;=\; (\psi^o,p)_X\,p\,.$$

But the condition (A5), the fact that $(\psi^o,p)_X = (\mathcal{K}\psi^o)(\hat{x}) \neq 0$, and $v^o \neq 0$ rule this out. Thus the multiplier λ is strictly positive.

Knowing then that $\lambda > 0$, we then go back to equation (7.74), multiply by a/v^o, set $\rho = \lambda a/v^o$, and obtain

$$\rho\big[I \;-\; c\mathcal{K}^*\mathcal{K}\big]\,\psi^o + \;\mathcal{K}^*\omega^2\mathcal{K}\psi^o \;=\; \frac{(\psi^o,p)_X}{v^o}\,p\,. \qquad (7.75)$$

We may, by introducing the function $\tilde{\psi}^o$ defined by

$$\tilde{\psi}^o \;:=\; \frac{v^o}{(\psi^o,p)_X}\,\psi^o\,,$$

reformulate the equation (7.75) together with the constraint $g(\tilde{\psi}^o) = 0$ as a non-linear system for $\tilde{\psi}^o$ and ρ

$$\rho\big[I \;-\; c\mathcal{K}^*\mathcal{K}\big]\tilde{\psi}^o + \;\mathcal{K}^*\omega^2\mathcal{K}\tilde{\psi}^o = p\,, \qquad (7.76a)$$

$$\Big(\tilde{\psi}^o, (I - c\mathcal{K}^*\mathcal{K})\tilde{\psi}^o\Big)_X = 0\,. \qquad (7.76b)$$

It is this last pair of equations that we use as the necessary conditions for the optimal solution and which we will treat numerically in the next two sections. Numerical treatment will require that we can compute the Riesz functional p. Fortunately, for specific situations the form of the Riesz functional p can be given concretely (see Example 7.15 below).

7.4.3 The Finite Dimensional Problems

Having established the existence of optimal solutions for the constrained *SNR*−problem as well as the necessary conditions for the infinite dimensional problem, we turn now to the question of finite-dimensional approximations to the optimal solutions. We follow the discussion in Section 3.2 where we introduced a general theory of approximation using the notion of a family of ultimately dense subspaces $\{X_n\}_{n=1}^\infty$, i.e. subspaces $X_n \subset X$ with $X_{n+1} \subset X_n$ for all n such that $\bigcup_n X_n$ is dense in X.

We assume that there exists an optimal solution ψ^o of $(\mathcal{P})$. Furthermore, we require that Assumptions (A5) and (A6) hold. Recall, that we have shown that the constraint is active at the optimal solution ψ^o, that is, $Q(\psi^o) = c$ (see Theorem 7.12).

We now replace the original problem $(\mathcal{P})$ by the problem $(\mathcal{P}_n)$ given by

$$(\mathcal{P}_n) \qquad \text{Maximize} \quad SNR(\psi) \quad \text{subject to} \quad \psi \in X_n \setminus \{0\}\,, \; Q(\psi) \;\leq\; c\,.$$

It is easy to show that this finite dimensional problem has a solution for n sufficiently large.

Theorem 7.13. *If there exists a strictly feasible input i.e., $Q(\varphi_0) < c$ for some $\varphi_0 \in X$ then there exists an integer n_0 such that the problem $(\mathcal{P}_n)$ has a solution $\psi_n^o \in X_n$ for every integer $n \geq n_0$.*

Proof: Since the family of finite dimensional subspaces $\{X_n\}_{n=1}^{\infty}$ is ultimately dense in X, and the functional $Q : X \to \mathbb{R}$ is continuous, there exists an integer n_0 and functions $\psi_n \in X_n$, $n \geq n_0$, with $Q(\psi_n) \leq c$. This means that, for n sufficiently large, there exists feasible points in the subspace X_n. As in the existence Theorem 7.10, we can use the homogeneity of the functionals *SNR* and Q to show that we can restrict ourselves to those $\psi \in X_n$ with $\|\psi\|_X = 1$, and the existence of optimal solutions for the problem $(\mathcal{P}_n)$ follows immediately. □

We may now show that the sequence of functions $\{\psi_n^o\}_{n=n_0}^{\infty}$, each element of which is an optimal solution of the corresponding problem $(\mathcal{P}_n)$, has accumulation points, each of which is an optimal solution of the original problem. We state this result precisely in the next theorem where we will denote the optimal values of the finite dimensional problems by v_n^o.

Theorem 7.14. *Let $\{\psi_n^o\}_{n=n_0}^{\infty}$ be a sequence of solutions to the problems $(\mathcal{P}_n)$ satisfying (without loss of generality) $\|\psi_n^o\|_X = 1$. Then this sequence has accumulation points and every such accumulation point is an optimal solution for the problem $(\mathcal{P})$.*

Proof: Again, we start with the weak compactness of the unit ball in the Hilbert space X. From this fact, we see that the sequence $\{\psi_n^o\}_{n=n_0}^{\infty}$ indeed has weak accumulation points. Let ψ^o be such an accumulation point and suppose that the subsequence $\{\psi_{n_\ell}\}_{\ell=1}^{\infty}$ converges weakly to ψ^o. Since the operator $\mathcal{K}$ is compact on X and the unit ball is closed in the weak topology, $\|\psi^o\|_X \leq 1$ and $\|\mathcal{K}(\psi_{n_\ell}^o - \psi^o)\|_{C(S^{d-1})} \to 0$ as $\ell \to \infty$. Indeed, ψ^o is admissible for $(\mathcal{P})$ since

$$Q(\psi^o) \;\leq\; \frac{1}{\|\mathcal{K}\psi^o\|_{L^2(S^{d-1})}^2} \;=\; \lim_{\ell\to\infty} \frac{1}{\|\mathcal{K}\psi_{n_\ell}^o\|_{L^2(S^{d-1})}^2} \;=\; \lim_{\ell\to\infty} Q(\psi_{n_\ell}^o) \;\leq\; c\,. \tag{7.77}$$

We now show that the accumulation point ψ^o is optimal for $(\mathcal{P})$. To do so, consider *any* optimal solution of $(\mathcal{P})$, say $\hat{\psi}$ with $\|\hat{\psi}\|_X = 1$. Then, since the constraint Q is active, we have $\|\mathcal{K}\hat{\psi}\|_{L^2(S^{d-1})}^2 = 1/c$. We construct an auxiliary sequence of strictly feasible functions $\hat{\psi}_\ell \in X_{n_\ell}$ such that $\hat{\psi}_\ell \to \hat{\psi}$. Indeed, for any $\varphi \in X$ we have for the derivative

$$\begin{aligned} Q'(\hat{\psi})\varphi &= \frac{2}{\|\mathcal{K}\hat{\psi}\|_{L^2(S^{d-1})}} \operatorname{Re}\left[\|\mathcal{K}\hat{\psi}\|_{L^2(S^{d-1})}\,(\varphi,\hat{\psi})_X \;-\; \|\hat{\psi}\|_X\,\left(\varphi,\mathcal{K}^*\mathcal{K}\hat{\psi}\right)_X\right] \\ &= \frac{2}{\|\mathcal{K}\hat{\psi}\|_{L^2(S^{d-1})}} \operatorname{Re}\left(\varphi\,,\,\frac{1}{c}\,\hat{\psi} - \mathcal{K}^*\mathcal{K}\hat{\psi}\right)_X . \end{aligned}$$

Therefore, under the assumption (A6) we can choose $\varphi \in X$ with $Q'(\hat{\psi})\varphi < 0$. By choosing $\|\varphi\|_X$ small enough we can assume that $\varphi_\ell^o := \hat{\psi} + \frac{1}{\ell}\varphi$ satisfy $Q(\varphi_\ell^o) < c$ for all $\ell = 1, 2, \ldots$

Now, since the family of finite dimensional subspaces is ultimately dense, for all $\ell = 1, 2, \ldots$ there exists $\hat{\psi}_\ell \in X_{n_\ell}$ with $\|\hat{\psi}_\ell - \varphi_\ell^o\|_X \le 1/\ell$. Since φ_ℓ^o is strictly feasible, $\hat{\psi}_\ell \in X_{n_\ell}$ can be chosen such that also $Q(\hat{\psi}_\ell) \le c$. Then $\hat{\psi}_\ell \in X_{n_\ell}$ are admissible for $(\mathcal{P}_{n_\ell})$ and $\hat{\psi}_\ell \to \hat{\psi}$ as ℓ tends to infinity.

Now, since ψ^o is admissible, we have

$$\begin{aligned} v^o \;\ge\; \mathit{SNR}(\psi^o) \;&=\; \lim_{\ell\to\infty} \mathit{SNR}(\psi_{n_\ell}^o) \;=\; \lim_{\ell\to\infty} v_{n_\ell}^o \\ \ge\; \lim_{\ell\to\infty} \mathit{SNR}(\hat{\psi}_\ell) \;&=\; \mathit{SNR}(\hat{\psi}) \;=\; v^o . \end{aligned}$$

Hence $\mathit{SNR}(\psi^o) = v^o$, and the accumulation point ψ^o is therefore optimal as claimed. Moreover $\lim_{\ell\to\infty} v_{n_\ell}^o = v^o$.

In fact, the constraint described by the functional Q is *active* at ψ^o as we have seen in Theorem 7.12 so that, since $\|\psi^o\|_X \le 1$, we have the estimate

$$\begin{aligned} 1 \ge \|\psi^o\|_X^2 \;&=\; c\,\|\mathcal{K}\psi^o\|_{L^2(S^{d-1})}^2 \;=\; c\lim_{\ell\to\infty} \|\mathcal{K}\psi_{n_\ell}^o\|_{L^2(S^{d-1})}^2 \\ &\ge \lim_{\ell\to\infty} \|\psi_{n_\ell}^o\|_X^2 \;=\; 1\,, \end{aligned}$$

hence $\|\psi^o\|_X = 1$. Moreover, we can show that the sequence $\psi_{n_\ell}^o \to \psi^o$ converges not only weakly in X but even *strongly*. Indeed, using the binomial theorem,

$$\|\psi_{n_\ell}^o - \psi^o\|_X^2 \;=\; \|\psi_{n_\ell}^o\|_X^2 + \|\psi^o\|_X^2 - 2\operatorname{Re}\left(\psi_{n_\ell}^o, \psi^o\right)_X \;\longrightarrow\; 2 - 2\,\|\psi^o\|_X^2 \;=\; 0 .$$

This completes the proof. □

We can now apply the Lagrange multiplier rule to the optimization problem $(\mathcal{P})$. We assume from now on that the sequence $\{\psi_n^o\}$ of solutions of $(\mathcal{P}_n)$ converge to a solution ψ^o of $(\mathcal{P})$. We introduce the orthogonal projection operator $P_n : X \to X_n$ by $(\varphi, \psi_n)_X = (P_n\varphi, \psi_n)_X$ for all $\varphi \in X$ and $\psi \in X_n$. First, we have to check the generalized Slater condition (7.72) holds for the finite dimensional problems. We have for $\psi_n \in X_n$

$$\begin{aligned} g(\psi_n^o) \;+\; g'(\psi_n^o)\psi_n \le g'(\psi_n^o)\psi_n \;&=\; 2\,\mathrm{Re}\left(\psi_n^o - c\,\mathcal{K}^*\mathcal{K}\psi_n^o\,,\,\psi_n\right)_X \\ &= 2\,\mathrm{Re}\left(\psi_n^o - c\,P_n\mathcal{K}^*\mathcal{K}\psi_n^o\,,\,\psi_n\right)_X . \end{aligned}$$

Under the assumption (A6) there exists $\psi \in X$ such that $\mathrm{Re}\left(\psi^o - c\,\mathcal{K}^*\mathcal{K}\psi^o\,,\,\psi\right)_X < 0$. With $\psi_n = P_n\psi$ we have

$$\begin{aligned} g(\psi_n^o) \;+\; g'(\psi_n^o)\psi_n &\le 2\,\mathrm{Re}\left(\psi_n^o - c\,P_n\mathcal{K}^*\mathcal{K}\psi_n^o\,,\,\psi_n\right)_X \\ &= 2\,\mathrm{Re}\left(\psi_n^o - c\,P_n\mathcal{K}^*\mathcal{K}\psi_n^o\,,\,\psi\right)_X \\ &= 2\,\mathrm{Re}\left(\psi^o - c\,\mathcal{K}^*\mathcal{K}\psi^o\,,\,\psi\right)_X \;+\; 2\,\mathrm{Re}\left(\psi_n^o - \psi^o, \psi\right)_X \\ &\quad - 2c\,\mathrm{Re}\left(P_n\mathcal{K}^*\mathcal{K}\psi_n^o - \mathcal{K}^*\mathcal{K}\psi^o\,,\,\psi\right)_X . \end{aligned}$$

Now we observe that the last two terms tend to zero as n tends to infinity. Therefore, the generalized Slater condition (7.72) is satisfied for sufficiently large n.

Noting that the adjoint of $\mathcal{K}_n = \mathcal{K}|_{X_n}$ is given by $P_n\mathcal{K}^*$ application of the multiplier rule just as in the derivation of (7.76a), (7.76b) yields the existence of $\rho_n \ge 0$ with

$$\rho_n\left[I - c\,P_n\mathcal{K}^*\mathcal{K}\right]\tilde{\psi}_n^o \;+\; P_n\mathcal{K}^*\omega^2\mathcal{K}\tilde{\psi}_n^o = P_n p\,, \tag{7.78a}$$

$$\left(\tilde{\psi}_n^o, (I - c\,P_n\mathcal{K}^*\mathcal{K})\tilde{\psi}_n^o\right)_X = 0\,. \tag{7.78b}$$

We note, however, that the discrete form of (A5), i.e.,

$$P_n\mathcal{K}^*(\omega^2\mathcal{K}\psi_n^o) \;\notin\; \mathrm{span}\,\{P_n p\}\,,$$

is never satisfied. Indeed, the operator $P_n\mathcal{K}^*\omega^2\mathcal{K}|_{X_n}$ from the finite dimensional space X_n into itself is one-to-one (since $P_n\mathcal{K}^*(\omega^2\mathcal{K}\psi_n) = 0$ implies that $\|\omega\mathcal{K}\psi_n\|_{L^2(S^{d-1})} = 0$ and thus $\psi_n = 0$) and therefore onto (i.e. *surjective*).

Therefore, we can not guarantee that ψ_n^o is active or $\rho_n > 0$. Nevertheless, we can proceed as in the original problem: If $\rho_n > 0$ then the optimal solution $\psi_n^o \in U_n$ satisfies the system (7.78a), (7.78b). If $\rho_n = 0$ then $\psi_n^o \in U_n$ satisfies the equation

$$P_n\mathcal{K}^*(\omega^2\mathcal{K}\psi_n^o) \;=\; P_n p\,. \tag{7.78c}$$

For the solution of this equation we have to check that the constraint $Q(\psi_n^o) \le c$ is satisfied.

We want to illustrate the use of the multiplier equation for the case where the operator $\mathcal{K}$ is given by feeding a line source.

Example 7.15. For the linear line source we have seen in Section 4.5, and likewise in Subsection 7.3.2, that

$$(K\psi)(t) \;=\; \alpha(t)\int_{-\ell}^{\ell} \psi(s)\,\mathrm{e}^{-ikts}\,ds\,, \quad |t| \le 1,$$

where $\alpha \in C[-1,1]$ is positive on $(-1,1)$. We think of $\alpha \equiv 1$ or $\alpha(t) = \sqrt{1-t^2}$. Therefore, $(K\psi)(\tau_0) = (\psi, p)_{L^2(-\ell,\ell)}$ with $p(s) = \alpha(\tau_0)\exp(ik\tau_0 s)$, $|s| \leq \ell$. In Section 4.5 we have computed

$$(K^*K\psi)(t) \;=\; \int_{-\ell}^{\ell} a(t-s)\,\psi(s)\,ds\,, \quad |t| \leq \ell\,,$$

where

$$a(\tau) \;=\; \int_{-1}^{1} \alpha(s)^2\, \mathrm{e}^{iks\tau}\, ds\,, \quad \tau \in \mathbb{R}\,,$$

and, analogously,

$$(K^*\omega^2 K\psi)(t) \;=\; \int_{-\ell}^{\ell} b(t-s)\,\psi(s)\,ds\,, \quad |t| \leq \ell\,,$$

with

$$b(\tau) \;=\; \int_{-1}^{1} \omega(s)^2\,\alpha(s)^2\, \mathrm{e}^{iks\tau}\, ds\,, \quad \tau \in \mathbb{R}\,.$$

Therefore, equation (7.76a) takes the form

$$\rho\,\psi^o(t) - \rho\, c \int_{-\ell}^{\ell} a(t-s)\,\psi^o(s)\,ds + \int_{-\ell}^{\ell} b(t-s)\,\psi^o(s)\,ds \;=\; \alpha(\tau_0)\,\mathrm{e}^{ikt\,\tau_0}\,, \quad (7.79)$$

for $|t| \leq \ell$ where we wrote ψ^o for $\tilde{\psi}^o$.

For every $\rho > 0$ this is a Fredholm integral equation of the second kind. If the homogeneous system has only the trivial solution $\psi \equiv 0$ then (7.79) has a unique solution $\psi^o = \psi_\rho$[4]. We determine $\rho > 0$ such that ψ_ρ satisfies also (7.76b), i.e.,

$$F(\rho) \;:=\; \big(\psi_\rho\,,\,(I - cK^*K)\psi_\rho\big)_{L^2(-\ell,\ell)} \;=\; 0\,.$$

To solve the equation (7.79) we rely on a numerical procedure and so need to make a finite dimensional approximation. We normalize and set $\ell = 1$. For the subspace X_n we take the space of all (algebraic) polynomials of order at most n, i.e. $X_n = \mathcal{P}_n$. As an orthonormal basis in X_n we take the Legendre polynomials L_j, $j = 0, \ldots, n$, normalized to $\|L_j\|_{L^2(-1,1)} = 1$.

We now derive the projected equation corresponding to (7.79). We make the ansatz

[4] This result is part of the Riesz Theory (see Theorem A.40)

$$\psi_n^o \;=\; \sum_{j=0}^{n} x_j\, L_j \quad \text{for some } \alpha_j \in \mathbb{C}\,,$$

replace ψ^o by ψ_n^o in (7.79), multiply (7.79) by $L_k(t)$ and integrate. This yields the approximate equation

$$\rho \sum_{j=0}^{n} x_j \int_{-1}^{1} L_j(t)\, L_k(t)\, dt - \sum_{j=0}^{n} x_j \int_{-1}^{1}\int_{-1}^{1} \big[\rho\, c\, a(t-s) - b(t-s)\big]\, L_j(s)\, L_k(t)\, ds\, dt$$
$$= \alpha(\tau_0) \int_{-1}^{1} \mathrm{e}^{ikt\tau_0}\, L_k(t)\, dt\,.$$

This equation holds for every $k = 0, \ldots, n$. Using the orthogonality of $\{L_k\}$ and the abbreviations

$$A_{km} := \int_{-1}^{1}\int_{-1}^{1} a(t-s)\, L_m(s)\, L_k(t)\, ds\, dt\,,$$
$$B_{km} := \int_{-1}^{1}\int_{-1}^{1} b(t-s)\, L_m(s)\, L_k(t)\, ds\, dt\,, \quad y_k \;:=\; \alpha(\tau_0) \int_{-1}^{1} \mathrm{e}^{ikt\tau_0}\, L_k(t)\, dt$$

for $k, m = 0, \ldots, n$, we may rewrite this last equation as

$$\rho\,(I - cA)x \;+\; Bx \;=\; y \quad \text{where } x = (x_j)_{j=0}^n \text{ and } y = (y_j)_{j=0}^n\,. \tag{7.80}$$

Although this is a finite linear system, the vector $y \in \mathbb{C}^{n+1}$ and the matrices $A, B \in \mathbb{C}^{(n+1)\times(n+1)}$ still have entries which must be evaluated numerically by quadrature formulae. It is convenient to take the Gauss-Legendre formula, i.e. replace y_k and A_{km}, B_{km} by

$$y_k \approx \alpha(\tau_0) \sum_{j=0}^{n} w_j\, \mathrm{e}^{ikt_j\tau_0}\, L_k(t_j)\,, \quad j = 0, \ldots, n\,,$$
$$A_{km} \approx \sum_{i,j=0}^{n} w_i\, w_j\, a(t_i - t_j)\, L_m(t_j)\, L_k(t_i)\,, \quad k, m = 0, \ldots, n\,,$$

and analogously for B. Here, $t_i \in (-1, 1)$ and $w_i \in \mathbb{R}$ are the Gauss-Legendre nodes and weights, respectively. We define the matrix $\mathbb{P} \in \mathbb{R}^{(n+1)\times(n+1)}$ by $\mathbb{P}_{mj} := \sqrt{w_j}\, L_m(t_j)$ and note that $\mathbb{P}$ is orthogonal since

$$\big(\mathbb{P}\,\mathbb{P}^\top\big)_{km} \;=\; \sum_{j=0}^{n} w_j\, L_m(t_j)\, L_k(t_j) \;=\; \int_{-1}^{1} L_m(t)\, L_k(t)\, dt \;=\; \delta_{km}\,.$$

With $\tilde{r}_j = \sqrt{w_j}\,\alpha(\tau_0)\exp(ikt_j\tau_0)$ and $\tilde{A}_{ij} = \sqrt{w_i w_j}\,a(t_i - t_j)$ and $\tilde{B}_{ij} = \sqrt{w_i w_j}\,b(t_i - t_j)$ we write

$$\sum_{i=0}^{n} w_i\, r(t_i)\, L_k(t_i) \;=\; (\mathbb{P}\tilde{r})_k \quad\text{and}\quad \sum_{i,j=0}^{n} w_i\, w_j\, a(t_i - t_j)\, L_m(t_j)\, L_k(t_i) \;=\; (\mathbb{P}\tilde{A}\mathbb{P}^\top)_{km}$$

and analogously for B.

Therefore, the equation (7.80) leads to the matrix equation

$$\rho\big(I - c\mathbb{P}\tilde{A}\mathbb{P}^\top\big)\hat{x} \;+\; \mathbb{P}\tilde{B}\mathbb{P}^\top\,\hat{x} \;=\; \mathbb{P}\,\tilde{r}$$

where $\hat{x} \in \mathbb{C}^{(n+1)\times(n+1)}$ is the approximation of $x \in \mathbb{C}^{(n+1)\times(n+1)}$ due to the replacements of y_k, A_{km}, and B_{km} by the Gauss-Legendre formulae. Finally, we set

$$\tilde{x} \;:=\; \mathbb{P}^\top\,\hat{x}$$

and note that $\mathbb{P}\,\tilde{x} = \hat{x}$. This yields

$$\rho\big(\mathbb{P} - c\mathbb{P}\tilde{A}\big)\,\tilde{x} \;+\; \mathbb{P}\tilde{B}\,\tilde{x} \;=\; \mathbb{P}\,\tilde{r}$$

or, since $\mathbb{P}$ has full rank,

$$\rho\big(I - c\tilde{A}\big)\tilde{x} \;+\; \tilde{B}\,\tilde{x} \;=\; \tilde{r}\,. \tag{7.81}$$

The approximate solution is then given by

$$\psi_n^o \;=\; \sum_{j=0}^{n} x_j\, L_j \;=\; \sum_{j=0}^{n} \big(\mathbb{P}\tilde{x}\big)_j\, L_j\,.$$

We compare this system with the one which has been derived by using the Nyström method for solving (7.79) (see Subsection 4.5.3 for the synthesis problem). The matrix $\tilde{B}$ and the right hand side coincide with the ones from the Nyström method. For the approximate solution ψ_n^o we have

$$\sqrt{w_k}\,\psi_n^o(t_k) \;\approx\; \sum_{j=0}^{n} \sqrt{w_k}\,L_j(t_k)\,\big(\mathbb{P}\,\tilde{x}\big)_j \;=\; \big(\mathbb{P}^\top\,\mathbb{P}\,\tilde{x}\big)_k \;=\; \tilde{x}_k$$

which corresponds to the solution of (4.58). Therefore, the projection and the Nyström method coincide for this particular choice of basis function and numerical evaluation of the matrix elements.

For the following example we have taken $\alpha \equiv 1$, $\ell = 1$, ω to be the characteristic function of $[0, 60^o] \cup [120^o, 180^o]$, $\tau_0 = \cos(90^o) = 0$, $\lambda = 1$ and $\lambda = \pi$, and some values of $c > 0$. The optimal signal-to-noise ratios and the corresponding values of ρ are listed in the following table.

	$\lambda = 1$		$\lambda = \pi$	
c	v^o	ρ	v^o	ρ
1	$1/\,\|A\| > 1!$		5.552	$1.4230e-1$
1.1	456	$1.1586e-2$	5.985	$1.2285e-1$
2	1212	$1.1969e-3$	9.378	$3.8238e-2$
3	1907	$2.8354e-4$	10.973	$5.3319e-3$
5	2262	$3.8725e-5$	11.528	$1.6996e-3$
10	2576	$1.7310e-5$	12.209	$8.9858e-4$
15	2805	$1.2331e-5$	12.693	$6.7294e-4$
20	3008	$9.8453e-6$	13.010	$5.5562e-4$

The following plots show the far field patterns $|K\psi^o|^2$ for $c = 3$ and wave lengths $\lambda = 1$ and $\lambda = \pi$, respectively. We have normalized ψ^o such that $\|\psi^o\|_{L^2(-1,1)} = 1$.

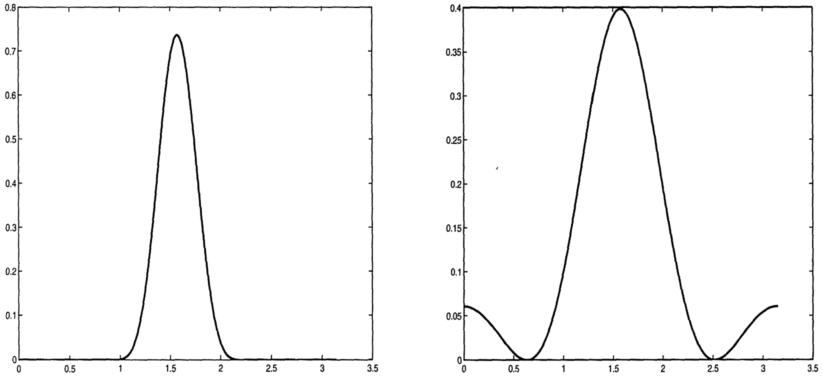


Fig. 7.8. Plots of $|K\psi^o|^2$ for wave lengths $\lambda = 1$ and $\lambda = \pi$, respectively, and $c = 3$

8

Conflicting Objectives: The Vector Optimization Problem

8.1 Introduction

In Chapter 3 we developed a coherent approach to the problem of determining optimal antenna feedings. By formulating various measures of performance as real-valued functionals defined in appropriate function spaces, we systematically used the methods of functional analysis and optimization theory. This approach allows us to study the existence and properties of optimal solutions as well as computational procedures for the numerical approximation of these solutions in particular cases. We provided examples of such concrete analyses, including computational results in Chapters 4 and 7.

The analysis presented in those chapters is restricted to problems for which there is a single design criterion or cost functional. However typical problems which arise in antenna design require us to consider, simultaneously, several, often conflicting, goals. It has long been recognized, for example, that the narrow focusing of the main beam of an antenna has the concomitant effect of increasing near-field power or the side-lobe level. In his classic 1946 paper [34], Dolph explicitly mentions such trade-offs:

> "In many applications it is more important to sacrifice some gain and beam width in order to achieve low-level side lobes. Several schemes have been suggested as a means of accomplishing this."

He proceeds to describe the "binomial feeding" which reduces the side lobes, as well as the method introduced by Schelkunoff [118]. Dolph viewed his own contribution as a "third means of improving the pattern of linear arrays".

The approach used most often in such situations is to select one of the several performance measures as the primary goal to be optimized, and set arbitrary but practical levels for the others, thereby introducing equality or inequality constraints. The problem is then treated as an ordinary constrained minimization or maximization problem. The treatment of constrained optimization problems was illustrated by the examples in Chapters 4 and 7. In one case, we introduced a constraint on the power available to the antenna by considering surface currents which are bounded in some appropriate norm; in another, we required the so-called quality factor i.e. the ratio of input power to far field power, both measured relative to perhaps different function-space norms, to be bounded.

In this chapter we take another approach to antenna design problems: the approach of multi-criteria optimization. Multi-criteria optimization concerns the *simultaneous* optimization of several distinct criteria over a set of inputs which are perhaps subject to prescribed constraints. Without loss of generality we assume that all criteria are to be minimized. In such problems, we search for so-called *Pareto points*, namely points for which any perturbation of their components causes at least one of the several objective functionals to increase. We will make this rough definition precise in the next section.

While well known in other applied fields, these techniques have only recently been applied to problems of antenna design [6], [12], [141], and [140]. We have also compared the application of these techniques to the results obtained from the single criterion treatment of the null-placement problem in [7]. The subject of multi-criteria optimization has been most thoroughly developed in the literature of mathematical economics and is most often associated there with the names of Walras and Pareto. It was the latter who introduced the basic notions in the late 1890's. The interested reader may consult the review article of Stadler [126] for historical background and the article of Dauer and Stadler [31] for more recent developments. Applications to problems in mechanical engineering are described in [127] which has an extensive bibliography. A more complete mathematical reference is the book of Jahn [56].

For those readers who are not familiar with the basic ideas of vector optimization we refer to the last section of the mathematical appendix where we present the necessary background material including general conditions insuring the existence of Pareto points and necessary conditions in the form of a multiplier rule. Section 8.2 will be dedicated to the case of arrays and, in particular, an analysis of the Dolph problem from the point of view of vector optimization following the lines of [13]. Section 8.3 studies the simultaneous maximization of power in two angular sectors for the case of the linear line source. Section 8.4 contains an analysis and numerical results for the signal-to-noise ratio problem first treated as a multi-criteria optimization problem in [6]. The multi-criterion problem for optimal power has been investigated computationally in [60].

8.2 General Multi-criteria Optimization Problems

In previous chapters we have dealt with optimization of real-valued cost functions. The present chapter is devoted to multi-criteria optimization problems and we discuss this general optimization problem here in order to prepare for the concrete examples which we study in this chapter just as we did with Section 3.2 of Chapter 3.

The general form of the **multi-criteria optimization problem** can be stated in terms of a vector-valued function $\mathcal{F}$. Specifically, given a Hilbert space X over $\mathbb{F}$, a subset $U \subset X$, an ordered Banach space Z, and a map

$$\mathcal{F} : U \longrightarrow Z, \tag{8.1}$$

the optimization problem is to find $\psi^o \in U$, such that

$$\mathcal{F}(\psi^o) \;<_\Lambda\; \mathcal{F}(\psi) \quad \text{for all } \psi \in U. \tag{8.2}$$

Since the values $\mathcal{F}(\psi) \in Z$, it is necessary to interpret the inequality sign in the relation (8.2). This inequality is defined in terms of the **order cone** Λ which defines a **partial order** in the ordered vector space Z. All the pertinent definitions and results leading up to this formulation may be found in Section A.9 of the Appendix which the reader should consult. In particular, (8.2) is the short form for $\mathcal{F}(\psi) - \mathcal{F}(\psi^o) \in \Lambda$ for all $\psi \in U$.

We can illustrate the idea by looking at the following example

Example 8.1. Let $X = \mathbb{R}^2$, $U \subset X$ be the unit disk, and take $Z = \mathbb{R}^2$ ordered by the **standard order cone** $\Lambda = \{\boldsymbol{x} \in \mathbb{R}^2 : x_1 \geq 0\,,\; x_2 \geq 0\}$ (see Definition A.25 in the Appendix). Define $\boldsymbol{\mathcal{F}} : U \to \mathbb{R}^2$ by $\boldsymbol{\mathcal{F}}(\boldsymbol{x}) := (x_1 - 3, x_2 - 3)^\top$. Then $\boldsymbol{\mathcal{F}}$ maps the disk U onto $B[\boldsymbol{z}_o, 1] = \{\boldsymbol{z} \in \mathbb{R}^2 : \|\boldsymbol{z} - \boldsymbol{z}_o\| \leq 1\}$ where $\boldsymbol{z}_o := (3,3)^\top$ by a simple translation and we are looking for all $\boldsymbol{x}^o \in U$ such that $\boldsymbol{\mathcal{F}}(\boldsymbol{x}^o) <_\Lambda \boldsymbol{\mathcal{F}}(\boldsymbol{y})$ for all $\boldsymbol{y} \in U$. A simple sketch will convince the reader that these solutions of the multi-criteria problem are just the points on the boundary of U which lie in the third quadrant (see the Figure 8.1 and Example 8.3 below).

This example illustrates the basic idea which we will make precise in a moment: our objective is to use the ordering in the range of a vector-valued function $\boldsymbol{\mathcal{F}} : U \to \mathbb{R}^2$, to find points which are desirable in the sense that, by a choice of a different input from U, we cannot *simultaneously* improve both the components of the vector criterion. We shall call points with this property, **minimal points** of the range of $\boldsymbol{\mathcal{F}}$; the *pre-image* of a minimal point is what is usually called a **Pareto point** for $\boldsymbol{\mathcal{F}}$.

In the applications that we consider in this chapter, we will need to work in a much broader setting and now turn to the appropriate generalization.

8.2.1 Minimal Elements and Pareto Points

In problems of vector optimization we are interested in *minimal elements* relative to a given order cone. The needed preliminary definitions may be found in Section A.9 of the Appendix.

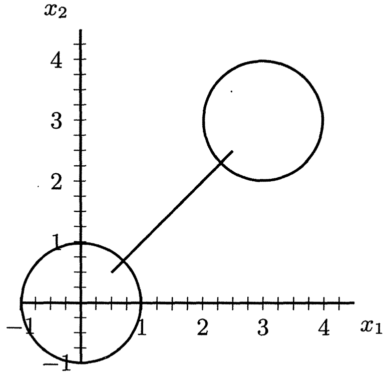


Fig. 8.1. The Vector Map $\mathcal{F}$

Definition 8.2. *Let $S \neq \emptyset$ be a subset of an ordered real vector space Z. Then $x^o \in S$ is a* **minimal element** *of S provided $x \in S$ and $x <_\Lambda x^o$ implies $x = x^o$.*

We can write this briefly as $(x^o - \Lambda) \cap S = \{x^o\}$ where we use the obvious notation

$$x^o - \Lambda \;=\; \{x^o - x : x \in \Lambda\}\,. \tag{8.3}$$

We mention two examples here. We first return to Example 8.1:

Example 8.3. In the case of the unit disk U, we see that that minimal points with respect to the standard order cone in $\mathbb{R}^2$ are just the points on the boundary that lie in the third quadrant. Indeed, if we have a point $\boldsymbol{x} = (x_1, x_2)^\top \in \mathbb{R}^2$, with $x_1 \leq 0, x_2 \leq 0$, and $\|\boldsymbol{x}\| = 1$ then, for any $\boldsymbol{y} = (y_1, y_2)^\top \in \mathbb{R}^2$ with $y_1 < x_1 \leq 0$ and $y_2 < x_2 \leq 0$ we have $|x_i| < |y_i|$, $i = 1, 2$, so that $\|\boldsymbol{y}\| \geq 1$ and therefore $\boldsymbol{y} \notin U$. Hence all such points $\boldsymbol{x}$ in the third quadrant and on the boundary of U are minimal.

The previous example is one in which the set of minimal points is bounded. However this is not necessarily the case. Let us look at the example of the region in the first quadrant of $\mathbb{R}^2$ lying above the graph of the hyperbola $x_1 x_2 = 1$.

Example 8.4. Let $Z = \mathbb{R}^2$ with $\Lambda = \big\{(x_1, x_2) \in \mathbb{R}^2 : x_1 \geq 0,\ x_2 \geq 0\big\}$ as order cone. Let $S = \big\{(x_1, x_2) \in \mathbb{R}^2 : x_1 \geq 0,\ x_2 \geq 0,\ x_1 x_2 \geq 1\big\}$. Then all points of the set

$$\big\{(x_1, x_2) \in S : x_1 x_2 = 1\big\}$$

are minimal and the set of minimal points is unbounded.

The minimal points are always boundary points if the ordered vector space Z is equipped with a topology:

Theorem 8.5. *Let Z be an ordered real Banach space with order cone Λ and $S \subset Z$ be non-empty. Then the set, M, of minimal points of S is a subset of the boundary of S i.e., $M \subset \partial S$.*

Proof: Assume, on the contrary, that there exists $y \in M \setminus \partial S$. Let $z \in \Lambda$ be non-zero. Then, since $y \in \text{int}(S)$, there exists $\rho > 0$ with $y - \rho z \in S$. Furthermore, $\rho z \in \Lambda$. It follows that $y \neq y - \rho z \in S \cap (y - \Lambda)$. But since y is assumed to be minimal, we have, according to the definition, that $S \cap (y - \Lambda) = \{y\}$, which is a contradiction. □

In order to develop conditions guaranteeing that a set contains minimal points, we need to introduce the concept of a *polar cone.* To this end, let Z be a real Banach space with dual Z^*. Thus Z^* is the set of all continuous linear maps $z^* : Z \to \mathbb{R}$. Again, denote the action of an element $z^* \in Z^*$ on $z \in Z$ by $z^*(z)$. For an arbitrary set $S \subset Z$ we have

Definition 8.6. *Let $S \subset Z$. The* **polar** *of the set S is defined to be*

$$S^p \; := \; \left\{ z^* \in Z^* : z^*(z) \leq 0 \quad \textit{for all } z \in S \right\}. \tag{8.4}$$

Note that, by linearity of z^*, the set S^p is a convex cone in Z^* *regardless* of the nature of the set S. We will refer to S^p as the **polar cone** of S. We note that the polar cone Λ^p of the normal order cone $\Lambda = \{\boldsymbol{x} \in \mathbb{R}^n : x_j \geq 0,\ j = 1,2,\ldots,n\}$ is just given by $-\Lambda = \{\boldsymbol{x} \in \mathbb{R}^n : x_j \leq 0,\ j = 1,2,\ldots,n\}$.

We can now prove a theorem that guarantees the existence of minimal points under conditions of wide applicability.

Theorem 8.7. *Let Z be an ordered real Banach space with a non-trivial closed convex order cone Λ. Suppose that* $\text{int}(\Lambda^p) \neq \emptyset$. *Then every compact subset S of Z contains minimal points.*

Proof: Let $z^* \in \text{int}(\Lambda^p)$. Then $z^*(z) < 0$ for all $z \in \Lambda \setminus \{0\}$. Indeed, if $z^*(\tilde{z}) = 0$ for some $\tilde{z} \in \Lambda \setminus \{0\}$ then, by the Hahn-Banach Theorem (see A.41) there exists $\tilde{z}^* \in Z^*$ with $\tilde{z}^*(\tilde{z}) = 1$ and thus $(z^* + \epsilon \tilde{z}^*)(\tilde{z}) = \epsilon > 0$. But $z^* + \epsilon \tilde{z}^* \in \Lambda^p$ for sufficiently small $\epsilon > 0$, a contradiction. Therefore, $z^*(z) < 0$ for all $z \in \Lambda \setminus \{0\}$. In particular, $z^* \neq 0$.

Now, for the given z^*, consider the map

$$z \mapsto z^*(z), \quad z \in S,$$

of $S \to \mathbb{R}$. By continuity and compactness this map has a maximum on S, say at $z^o \in S$, i.e.,

$$z^*(z^o) \; \geq \; z^*(z) \quad \text{for all} \quad z \in S. \tag{8.5}$$

Now assume that there exists $\hat{z} \in S$ with $\hat{z} \neq z^o$ and $\hat{z} <_\Lambda z^o$. Then $z^o - \hat{z} \in \Lambda \setminus \{0\}$ and thus $z^*(z^o - \hat{z}) < 0$, a contradiction to (8.5). □

Remark: In the applications that we have discussed earlier in this book, the space $Z = \mathbb{R}^n$ and the usual order cone indeed satisfies the conditions of this theorem while the polar cone likewise has a non-empty interior.

It may well happen, particularly in the case when the set S is the range of some mapping into Z, that the set is neither closed nor bounded. In such a case, we need a weaker condition.

Theorem 8.8. *Let Z be an ordered real Banach space with a non-trivial closed convex order cone Λ. Suppose that* $\mathrm{int}(\Lambda^p) \neq \emptyset$ *and that $S \subset Z$. Then if, for some $z \in Z$,*

$$S_z \ = \ (z - \Lambda) \ \cap \ (S + \Lambda) \ = \ \left\{ s + y : s \in S, \ y \in \Lambda, \ s + y <_\Lambda z \right\}, \quad (8.6)$$

is nonempty and compact, then S contains a minimal point.

Proof: Since, by hypothesis, S_z is compact, it contains a minimal point by the previous theorem. Suppose that z^o is such a minimal point for S_z. Then, by definition,

$$(z^o - \Lambda) \ \cap \ S_z \ = \ \{z^o\} . \quad (8.7)$$

We have to show that $z^o \in S$ and $(z^o - \Lambda) \cap S = \{z^o\}$. Since $z^o \in S_z$ we have $z^o \in \hat{z} + \Lambda$ for some $\hat{z} \in S$. But then, $\hat{z} \in z^o - \Lambda \subset z - \Lambda$ and $\hat{z} \in S_z$. By (8.7), $z^o = \hat{z} \in S$ follows. We may take now any $x \in S \cap (z^o - \Lambda)$. Then $x \in z^o - \Lambda$ and also $z^o \in z - \Lambda$ since $z^o \in S_z$. By adding these inclusions we get $x \in z - \Lambda$, i.e. $x \in (z^o - \Lambda) \cap S_z = \{z^o\}$. Thus z^o is also minimal for S. □

The general multi-criteria optimization problem can now be formulated as follows:

> Given a linear space[1] X and an ordered real Banach space Z, let $U \subset X$ and suppose $\mathcal{F} : U \to Z$. Find $x^o \in U$ such that $\mathcal{F}(x^o)$ is a minimal element of $\mathcal{F}(U)$.

We can now give a precise definition of what is meant by a Pareto optimal point relative to a specific vector-valued criteria function, $\mathcal{F} : U \to Z$ for $U \subset X$, a vector space, and Z an ordered real Banach space.

Definition 8.9. *A point $x^o \in U$ is said to be* **Pareto optimal** *relative to the vector-valued function $\mathcal{F}$ provided $\mathcal{F}(x^o)$ is minimal with respect to $\mathcal{F}(U)$.*

The term Pareto optimal is chosen here for historical reasons. Other terms have been used frequently including "non-inferior solutions", "non-dominated

[1] In our applications, X will be a separable Hilbert space.

solutions," and "efficient solutions." Some of these terms may be more informative in that they better suggest the property which characterizes Pareto points, namely that we cannot lower one of the component values by moving from that point without strictly increasing at least one of the other components of the criteria vector. In general Pareto points are not unique since minimal points are not unique.

We remark that it is seldom the case that there exists some point that will minimize all components of the vector criterion simultaneously, nor is it necessarily true that standard scalar optimization methods can be used to find the Pareto set. In particular, it is not generally the case that the minimization of *one* criterion subject to inequality constraints on the others will yield a Pareto point.

Applications of Theorems 8.7 and 8.8 to $S = \mathcal{F}(U)$ immediately yield the following result.

Theorem 8.10. *Let X be a real or complex Banach space and Z a real ordered Banach space with a non-trivial closed convex order cone Λ. Suppose that $\mathrm{int}(\Lambda^p) \neq \emptyset$ and that $U \subset X$.*

(a) If $\mathcal{F}(U)$ is compact, then U contains Pareto points.
(b) If $\mathcal{F} : U \to Z$ is such that, for some $z \in Z$,

$$\begin{aligned} S_z &= (z - \Lambda) \cap \big(\mathcal{F}(U) + \Lambda\big) && (8.8)\\ &= \big\{\mathcal{F}(x) + y : x \in U,\ y \in \Lambda,\ \mathcal{F}(x) + y <_\Lambda z\big\} && (8.9) \end{aligned}$$

is nonempty and compact, then there exist Pareto points of $\mathcal{F}$ in U.

The compactness of $\mathcal{F}(U)$ often follows if $\mathcal{F}$ is completely continuous.

Definition 8.11. *Let X and Z be a pair of Banach spaces, and a set $U \subset X$. A map $\mathcal{F} : U \to Z$ is called* **completely continuous** *provided $\mathcal{F}$ maps weakly convergent sequences into norm convergent sequences.*

We note that if $Z = \mathbb{R}$ this notion coincides with the definition of weak sequential continuity of a map.

Theorem 8.12. *If X is a reflexive Banach space, $U \subset X$ a closed, bounded, and convex set, and $\mathcal{F}$ is completely continuous, then $\mathcal{F}(U)$ is compact in Z.*

Proof: Since X is reflexive, Alaoglu's Theorem (Theorem A.56) and Theorem A.58 show that the set U is weakly sequentially compact. Then $\mathcal{F}(U)$ is compact by the complete continuity of $\mathcal{F}$. □

We now wish to recall the ideas of a weak minimal point and a weak Pareto point. The concept of a weak minimal point and the corresponding notion of weak Pareto point are particularly important in studying the convergence approximation methods as we will explain presently.

Definition 8.13. *Let Z be an ordered real Banach space with order cone Λ having a non-empty interior. Then $z^o \in S \subset Z$ is called a* **weak minimal point** *of S provided*

$$\big(z^o - \mathrm{int}\,(\Lambda)\big) \,\cap\, S \;=\; \emptyset\,. \tag{8.10}$$

If X is a Banach space and $U \subset X$, then $x^o \in U$ is called a **weak Pareto point** *for $\mathcal{F} : U \to Z$ provided $\mathcal{F}(x^o)$ is a weak minimal point of the set $\mathcal{F}(U) \subset Z$ i.e.,*

$$\big(\mathcal{F}(x^o) - \mathrm{int}\,(\Lambda)\big) \,\cap\, \mathcal{F}(U) \;=\; \emptyset\,.$$

Remark: In the subsequent discussion we will often refer to Pareto points as **strong Pareto points** to distinguish them from weak Pareto points.

Let us look at a pair of examples.

Example 8.14.
We consider the space $\mathbb{R}^2$ with the standard order. The order cone is just $\Lambda = \big\{\boldsymbol{x} \in \mathbb{R}^2 : x_1 \geq 0,\; x_2 \geq 0\big\}$ which has interior $\mathrm{int}\,(\Lambda) = \big\{\boldsymbol{x} \in \mathbb{R}^2 : x_1 > 0,\; x_2 > 0\big\}$. Now let

$$S \;=\; \big\{\boldsymbol{x} \in \mathbb{R}^2 : 0 \leq x_1 \leq 1,\; x_1 \leq x_2 \leq 1\big\}\,.$$

We fix any point of the form $(0, x_2)$ with $0 \leq x_2 \leq 1$. Then the set

$$(0, x_2) \;-\; \mathrm{int}\,(\Lambda) \;=\; \big\{\boldsymbol{y} \in \mathbb{R}^2 : y_1 < 0,\; y_2 < x_2\big\}$$

does not meet S since, in particular, $y_1 < 0$. Hence the set $M = \big\{\boldsymbol{x} \in \mathbb{R}^2 : x_1 = 0,\; 0 \leq x_2 \leq 1\big\}$ is the set of weak minimal points. Note that, in this example, the only minimal point is $(0,0)$ as Figure 8.2 illustrates. (The shaded region in each represents a portion of $(0, x_2) - \Lambda$.)

The concept of a weak minimal point and the corresponding notion of weak Pareto point are particularly important in studying the convergence approximation methods. The hope is that by introducing finite dimensional approximation problems of sufficiently large dimension n, we obtain numerical solutions which approximate minimal solutions of the original problem. However, we can usually only guarantee that the solution of the finite dimensional problem is at most a weak Pareto point.

If we are in the situation described in Theorem 8.10 we can state the following approximation result.

Theorem 8.15. *Let U be a closed, bounded, and convex subset of a reflexive Banach space X. Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of subspaces of the Banach space X such that $\bigcup_n (U \cap X_n)$ is dense in U. Assume that $\mathcal{F} : U \to Z$ is completely continuous and let $x_n^o \in (U \cap X_n)$, $n = 1, 2, \ldots$, be a sequence of Pareto points for $\mathcal{F}$ on $U \cap X_n$. Then the sequence has weak accumulation points and every weak accumulation point of this sequence is a weak Pareto point for $\mathcal{F}$ on U.*

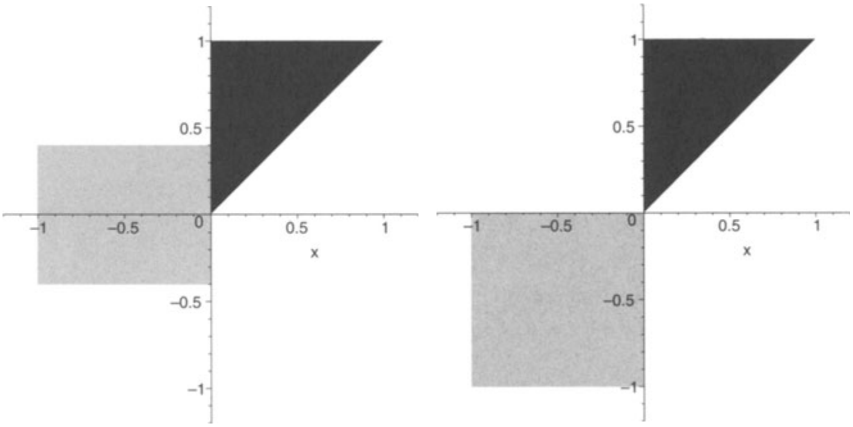


Fig. 8.2. Weak and Strong Minimal Points of S

Proof: Let $U_n = U \cap X_n$. Since U is weakly compact and $x_n^o \in U_n$, the sequence must have at least one weak accumulation point $x^o \in U$. Suppose this point is *not* a weak Pareto point. Then there must be a point $\hat{z} \in \mathcal{F}(U)$ such that $\mathcal{F}(x^o) - \hat{z} \in \operatorname{int}(\Lambda)$. Let $\hat{x} \in U$ be such that $\mathcal{F}(\hat{x}) = \hat{z}$. Since $\bigcup_n U_n$ is dense in U, it follows that there is a sequence $\{\tilde{x}_n\}_{n=1}^{\infty}$ with $\tilde{x}_n \in U_n$ for all n and $\hat{x} = \lim_{n\to\infty} \tilde{x}_n$. Since $\mathcal{F}$ is a completely continuous map,

$$\lim_{n\to\infty} \bigl(\mathcal{F}(x_n^0) - \mathcal{F}(\hat{x}_n)\bigr) \;=\; \bigl(\mathcal{F}(x^o) - \hat{z}\bigr) \;\in\; \operatorname{int}(\Lambda)\,.$$

Hence, for m a sufficiently large integer, $\mathcal{F}(x_m^o) - \mathcal{F}(\hat{x}_m) \in \operatorname{int}(\Lambda)$, and this contradicts the assumption that x_m^o is a minimal solution of the finite dimensional problem in U_m. Therefore, x^o is a weak minimal solution and the proof is complete. □

Remark: The assumption that $\bigcup_n (U \cap X_n)$ is dense in U is satisfied, e.g., if $\bigcup_n X_n$ is dense in X and the interior $\overset{o}{U}$ of U is non-empty, see Lemma 3.24.

8.2.2 The Lagrange Multiplier Rule

As it is well known from the optimization of scalar functions, the Lagrange multiplier rule often helps to compute the optimal solutions. We continue with a necessary condition in the form of a Lagrange multiplier rule for vector optimization problems and restrict ourselves to the special cases which are needed to treat the specific problems discussed in Section 8.3 and 8.4 below. A more general statement, as well as the proof, can be found in the book of Kirsch, Warth and Werner [71].

Theorem 8.16. *Let X be a Hilbert space satisfying the hypotheses of Theorem 8.10 and assume that $\mathbb{R}^n$ is ordered by the usual order cone. Let*

$\mathcal{F} : X \to \mathbb{R}^n$ be Fréchet differentiable while $g : X \to \mathbb{R}^m$ is continuously Fréchet differentiable. Let $U := \{x \in X : g_j(x) \le 0,\ j = 1, 2, \dots, m\}$. Suppose that $x^o \in U$ is a weak Pareto point for $\mathcal{F}$ on U. Then there exists exist $\eta = (\eta_1, \dots, \eta_n)^\top \in \mathbb{R}^n_{\geq 0}$ and $\rho = (\rho_1, \dots, \rho_m)^\top \in \mathbb{R}^m_{\geq 0}$, not vanishing simultaneously, such that

$$\rho_j g_j = 0 \textit{ for all } j = 1, \dots, m, \tag{8.11a}$$

and

$$\sum_{i=1}^{n} \eta_i \, \nabla F_i(x^o) \; + \; \sum_{j=1}^{m} \rho_j \, \nabla g_j(x^o) \; = \; 0 \, . \tag{8.11b}$$

Moreover, $(\eta_1, \dots, \eta_n) \neq (0, \dots, 0)$ if the following **constraint qualification** *is satisfied:*

$$\textit{The gradients} \quad \nabla g_j(x^o)\,, \; j = 1, \dots, m, \quad \textit{are linearly independent.} \tag{8.12}$$

We point out that in general the conditions (8.11) are only *necessary* for optimality. However, if all of the functions F_i and g_j are convex then they are also sufficient:

Theorem 8.17. *Let X be a Hilbert space satisfying the hypotheses of Theorem 8.10 and assume that $\mathbb{R}^n$ is ordered by the usual order cone. Let $\mathcal{F} : X \to \mathbb{R}^n$ and $g : X \to \mathbb{R}^m$ be Fréchet differentiable and let every component $F_i : X \to \mathbb{R}$ and $g_j : X \to \mathbb{R}$ be convex. Again set $U := \{x \in X : g_i(x) \le 0,\ i = 1, 2, \dots, m\}$. Suppose that there exists $x^o \in U$ and $\eta \in \mathbb{R}^n_{\geq 0}$ and $\rho \in \mathbb{R}^m_{\geq 0}$ such that (8.11) holds.*

If $\eta \neq 0$, i.e. at least one $\eta_i > 0$, then x^o is a weak Pareto point of $\mathcal{F}$ on U. If $\eta_i > 0$ for all $i = 1, \dots, n$, then x^o is a strong Pareto point.

Proof: It is well known that the convexity of a differentiable map F_i is equivalent to the inequality

$$F_i(x) \; - \; F_i(x^o) \; \geq \; \operatorname{Re}\big(\nabla F_i(x^o), x - x^o\big)_X\,, \quad i = 1, \dots, n,$$

for all $x \in X$ and, analogously,

$$g_j(x) \; - \; g_j(x^o) \; \geq \; \operatorname{Re}\big(\nabla g_j(x^o), x - x^o\big)_X\,, \quad j = 1, \dots, m,$$

for all $x \in X$.We multiply these inequalities by $\eta_i \ge 0$ and $\rho_j \ge 0$, respectively, and sum the resulting inequalities. This yields

$$\begin{aligned} &\sum_{i=1}^{n} \eta_i \big[F_i(x) - F_i(x^o)\big] \; + \; \sum_{j=1}^{m} \rho_j \big[g_j(x) - g_j(x^o)\big] \\ &\geq \left(\sum_{i=1}^{n} \eta_i \, \nabla F_i(x^o) \; + \; \sum_{j=1}^{m} \rho_j \, \nabla g_j(x^o)\,,\, x - x^o \right)_X \; = \; 0\,, \end{aligned}$$

i.e. for $x \in U$

$$\sum_{i=1}^{n} \eta_i \big[F_i(x) - F_i(x^o)\big] \;\geq\; -\sum_{j=1}^{m} \rho_j \big[g_j(x) - g_j(x^o)\big] \;=\; -\sum_{j=1}^{m} \rho_j \, g_j(x) \;\geq\; 0$$

since $\rho_j g_j(x^o) = 0$ for all j. This yields

$$\sum_{i=1}^{n} \eta_i \, F_i(x) \;\geq\; \sum_{i=1}^{n} \eta_i \, F_i(x^o) \quad \text{for all } x \in U\,.$$

Now let $\eta \neq 0$ and assume on the contrary that x^o is not a weak Pareto point of $\mathcal{F}$ on U. Then there exists $x \in U$ such that $F_i(x) < F_i(x^o)$ for all $i = 1, \ldots, n$. Since $\eta_i \geq 0$ for all i and strictly positive for some i we conclude that

$$\sum_{i=1}^{n} \eta_i \, F_i(x) \;<\; \sum_{i=1}^{n} \eta_i \, F_i(x^o)\,,$$

a contradiction.

Now let $\eta_i > 0$ for all i and assume, on the contrary, that x^o is not a strong Pareto point of $\mathcal{F}$ on U. Then there exists $x \in U$ such that $F_i(x) \leq F_i(x^o)$ for all $i = 1, \ldots, n$, and strict inequality holds for some i. This yields a contradiction as before. □

8.2.3 Scalarization

We remark that under the assumptions of the previous theorem all weak Pareto points are obtained by minimizing the *scalar* function

$$\mathcal{J}(x) \;:=\; \sum_{i=1}^{n} \eta_i \, F_i(x) \quad \text{over } U$$

for some $\eta_i \geq 0$ with $\sum_{i=1}^{n} \eta_i > 0$ and all strong Pareto points are obtained by minimizing $\mathcal{J}$ for some $\eta_i > 0$, $i = 1, \ldots, n$.

The term **scalarization** refers to the replacement of a multi-criteria optimization problem with one having a scalar, real-valued cost functional whose solutions are closely related to, if not identical with, the solutions of the original multi-criteria problem. Such a replacement is of great use for computational purposes since the algorithms for scalar optimization are highly developed.

We start again with a map $\mathcal{F}$ from a real or complex Banach space into a real ordered Banach space Z with a non-trivial closed convex order cone Λ. While there are many possible methods of scalarizing the minimization of $\mathcal{F}$ on some given set U of constraints, and we concentrate on one, **linear scalarization**[2],

[2] Other methods include, for example, *quadratic scalarization*, in which the multi-criteria objective function is replaced by a weighted L^2-norm of the components of $\mathcal{F}$ with positive weights. The scalar optimization problem in this case becomes one of minimal norm.

in which the scalar objective function is of the form $x \mapsto \ell\bigl(\mathcal{F}(x)\bigr)$ where ℓ is some element from the dual cone $\Lambda^* := -\Lambda^p$ of Λ. For example, in the case that $Z = \mathbb{R}^n$, the scalar objective is simply

$$\mathcal{J}(x) \;:=\; \sum_{i=1}^{n} \eta_i \, F_i(x)\,,$$

where it is usual to take $\eta_i \geq 0$, $i = 1, 2, \ldots, n$ with $\sum_{i=1}^{n} \eta_i = 1$. In this case $\eta \in \mathbb{R}^n$ represents a probability distribution over the components of the vector function $\mathcal{F}$.

On the one hand, it is true that we obtain Pareto points by minimizing such a scalarized problem. On the other, it may well be, as examples show, that there are Pareto points for the original problem which are not found by minimizing a particular scalarized problem. To do so, we need to know that additional conditions are satisfied. The situation is described more precisely in the following two results whose proofs may be found in the book of Jahn [56].

Theorem 8.18. *Let X and Z be Banach spaces and let $\mathcal{F} : X \longrightarrow Z$ where Z is ordered with order cone Λ. Let $\ell \in \Lambda^* \setminus \{0\}$. Then every point $x^o \in X$ which minimizes $x \mapsto \ell\bigl(\mathcal{F}(x)\bigr)$ on some set U is a Pareto point for the vector function $\mathcal{F}$ on U.*

Theorem 8.19. *Let X and Z be Banach spaces and let $\mathcal{F} : X \longrightarrow Z$ where Z is ordered with order cone Λ. Assume that x^o is a Pareto point for $\mathcal{F}$ on some set U. Then if the set $\mathcal{F}(U) + \Lambda$ is convex, there exists a $\ell \in \Lambda^* \setminus \{0\}$ such that x^o is a point at which the mapping $x \mapsto \ell\bigl(\mathcal{F}(x)\bigr)$ takes its absolute minimum.*

There are explicit conditions known which guarantee that $\mathcal{F}(U) + \Lambda$ is convex, which are given in Jahn [56].

As our applications involve a finite dimensional range space $Z = \mathbb{R}^n$, we use scalarization in the form of a Lagrange multiplier rule as set out in Theorem 8.16. The condition that the components, F_i of the vector cost function are each convex leads to the sufficiency result given in Theorem 8.10, so that the application of the latter theorem gives a test that a point be a Pareto point for the original multi-criteria problem.

8.3 The Multi-criteria Dolph Problem for Arrays

In this section we return to the problem of an array of $2n + 1$ uniformly spaced Co-linear radiating dipoles located symmetrically along the z-axis of a three dimensional Cartesian coordinate system. In Subsection 1.4.2 we analyzed the problem as a constrained optimization problem. Here, we confront the question raised by Dolph by adopting the viewpoint of multi-criteria optimization. In order to treat the problem analytically, we will find it convenient to reformulate it.

8.3.1 The Weak Dolph Problem

Recall from Chapter 1 that the radiation pattern of a dipole array is entirely determined by the magnitude of the **array factor**, $f(\theta)$,

$$f(\theta) \;=\; \sum_{n=-n}^{n} a_n \, \mathrm{e}^{-ikdn\cos\theta}\,, \quad 0 \le \theta \le \pi\,. \tag{8.13}$$

In the case of cophasal, symmetric excitations, i.e. $a_n = a_{-n}$, $n = 1, 2, \dots, n$, the a_n may be considered real, and the expression (8.13) for the array factor may be written in the form:

$$f(\theta) \;=\; a_0 \;+\; 2\sum_{n=1}^{n} a_n \cos(kdn\cos\theta)\,. \tag{8.14}$$

Moreover, since the array factor is symmetric with respect to $\theta = \frac{\pi}{2}$, i.e., $f(\pi/2+\theta) = f(\pi/2-\theta)$, $0 \le \theta \le \pi$, we may restrict consideration to the interval $[0, \pi/2]$.

Recalling that the set $\mathcal{T}_n$ is defined by

$$\mathcal{T}_n \;:=\; \left\{ a_0 + 2\sum_{n=1}^{n} a_n \cos(nkd\cos\theta) : a_n \in \mathbb{R} \right\}, \tag{8.15}$$

the single objective constrained **Dolph-Tschebyscheff Problem** is:

$$(\mathcal{P}_{DT}) \qquad \begin{array}{ll} \text{Minimize} & \max\limits_{0\le\theta\le\hat{\theta}} |f(\theta)| \\ \text{Subject to} & f \in \mathcal{T}_n\,,\; f(\hat{\theta}) = 0\,,\; f(\pi/2) = 1\,, \end{array}$$

where the main beam is to be 1 at $\pi/2$ and $\hat{\theta} < \pi/2$ determines the beamwidth.

In the present setting of multi-criteria optimization, we will find it helpful to modify this problem by introducing what we will call the **Weak Dolph Problem**, $(\mathcal{P}_W)$, for which we seek to minimize the *total* side-lobe power as follows:

For $\hat{\theta} \in (0, \pi/2)$ given,

$$(\mathcal{P}_W) \qquad \begin{array}{l} \text{Minimize} \quad \displaystyle\int_0^{\hat{\theta}} \mathrm{e}^{-\sigma(\hat{\theta}-\theta)}\, |f(\theta)|^2 \, d\theta \\ \text{Subject to } f \in \mathcal{T}_n\,,\; f(\hat{\theta}) = 0\,,\; f(\pi/2) = 1\,. \end{array}$$

Here, $\sigma \ge 0$ denotes a weight factor. In the case $\sigma = 0$ the integral is just the ordinary L^2-norm of the array factor on the interval $[0, \hat{\theta}]$. If $\sigma > 0$ portions of this interval close to $\hat{\theta}$ have stronger weights than those close to 0.

One obvious question is the relationship between the array patterns obtained by optimizing this new criterion and those of the original Dolph problem. To do this, we introduce the functions w_n, $n = 1, \dots, n$, by

$$w_n(\theta) \; := \; \cos(nkd \cos\theta) \; - \; \cos(nkd \cos\hat{\theta}). \tag{8.16}$$

Expanding $f \in \mathcal{T}_n$ in the form

$$f \; = \; \sum_{n=1}^{n} a_n \, w_n$$

guarantees that $f(\hat{\theta}) = 0$. Consequently, introducing the $n \times n$ symmetric matrix $\boldsymbol{W}$ whose entries are given by

$$W_{nm} \; := \; \int_0^{\hat{\theta}} w_n(\theta) \, w_m(\theta) \, \mathrm{e}^{-\sigma(\hat{\theta}-\theta)} \, d\theta \, , \tag{8.17}$$

the vector valued function

$$\boldsymbol{w}(\theta) \; := \; \big(w_1(\theta), \dots, w_n(\theta)\big)^\top \, , \tag{8.18}$$

and $\boldsymbol{b} = \boldsymbol{w}(\pi/2)$, we note that $(\mathcal{P}_{DT})$ and $(\mathcal{P}_W)$ can be written in the forms

$(\mathcal{P}_{DT})$ Minimize $\max\limits_{0 \le \theta \le \hat{\theta}} \left|\boldsymbol{a}^\top \boldsymbol{w}(\theta)\right|$ subject to $\boldsymbol{a} \in \mathbb{R}^n$, $\boldsymbol{a}^\top \boldsymbol{b} = 1$,

and

$(\mathcal{P}_W)$ Minimize $\boldsymbol{a}^\top \boldsymbol{W} \boldsymbol{a}$ subject to $\boldsymbol{a} \in \mathbb{R}^n$, $\boldsymbol{a}^\top \boldsymbol{b} = 1$,

respectively. Since $\boldsymbol{W}$ is positive definite, as can be easily checked, we are able to estimate the magnitudes of the side-lobes with respect to both of the norms. Indeed, using the Cauchy-Schwarz inequality with the Euclidean norm $\|\cdot\|_2$ we have

$$\max_{0 \le \theta \le \hat{\theta}} \left|\boldsymbol{a}^\top \boldsymbol{w}(\theta)\right|^2 \; \le \; \|a\|_2^2 \underbrace{\max_{0 \le \theta \le \hat{\theta}} \|\boldsymbol{w}(\theta)\|_2^2}_{=:c^2} \; \le \; \frac{c^2}{\lambda_{min}^2} \, \boldsymbol{a}^\top \boldsymbol{W} \boldsymbol{a}$$

where λ_{min} denotes the smallest eigenvalue of $\boldsymbol{W}$. The reverse estimate follows directly from

$$\begin{aligned} \boldsymbol{a}^\top \boldsymbol{W} \boldsymbol{a} &= \int_0^{\hat{\theta}} \mathrm{e}^{-\sigma(\hat{\theta}-\theta)} \, |f(\theta)|^2 \, d\theta \; \le \; \max_{0 \le \theta \le \hat{\theta}} |f(\theta)|^2 \int_0^{\hat{\theta}} \mathrm{e}^{-\sigma(\hat{\theta}-\theta)} \, d\theta \\ &= \max_{0 \le \theta \le \hat{\theta}} |f(\theta)|^2 \, \hat{\theta} \, \frac{1}{\sigma\hat{\theta}} \big[1 - \exp(-\sigma\hat{\theta})\big] \; \le \; \hat{\theta} \max_{0 \le \theta \le \hat{\theta}} |f(\theta)|^2 \\ &= \hat{\theta} \max_{0 \le \theta \le \hat{\theta}} \left|\boldsymbol{a}^\top \boldsymbol{w}(\theta)\right|^2 \end{aligned}$$

Here we used the inequality $\left[1 - \exp(-t)\right]/t \leq 1$ for all $t \geq 0$. Therefore, measuring the magnitude of the side-lobes in $(\mathcal{P}_{DT})$ and $(\mathcal{P}_W)$ is an equivalent task[3]. We call $(\mathcal{P}_W)$ the *weak* Dolph problem since the constant c/λ_{min} is very large (growing exponentially with respect to n) due to the smallness of the smallest eigenvalue λ_{min}.

Also, since $\boldsymbol{W}$ is symmetric and positive definite, the quadratic programming problem $(\mathcal{P}_W)$ has a **unique** optimal solution given by

$$\boldsymbol{a}^o = \left[\boldsymbol{b}^\top \boldsymbol{W}^{-1}\boldsymbol{b}\right]^{-1}\boldsymbol{W}^{-1}\boldsymbol{b} \tag{8.19}$$

with minimal value

$$\boldsymbol{a}^{o\top}\boldsymbol{W}\,\boldsymbol{a}^o = \left[\boldsymbol{b}^\top \boldsymbol{W}^{-1}\boldsymbol{b}\right]^{-1}. \tag{8.20}$$

(see Hestenes [47]).

Example 8.20. In this example we compare the array pattern for the optimal Dolph-Tschebyscheff solution f^o_{DT} of $(\mathcal{P}_{DT})$ with that of the optimal pattern f^o_W of $(\mathcal{P}_W)$ for three choices of the weight factor σ.

Figure 8.3 shows the results for $n = 3$, $\hat{\theta} = \frac{4}{5} \cdot \frac{\pi}{2}$ and inter-element spacings $d = \lambda/2$ (left) and $d = 2 \cdot \lambda/3$ (right). The solid line represent the (absolute value of the) classical Dolph-Tschebyscheff arrays f^o_{DT}, while the dashed line the optimal arrays f^o_W for problem $(\mathcal{P}_W)$ with weight factor $\sigma = 0$.

Comparison of the results for pointwise side-lobe constraints with the L^2 constraint shows that while in the case of L^2-minimization, the first side-lobe may be larger than the corresponding side-lobe for the problem $(\mathcal{P}_W)$. Subsequent side-lobes, are, however, lower. In Figures 8.4 and 8.5 we show the optimal patterns for weight factors $\sigma = 3$ and $\sigma = 5$, respectively, for the same configuration as before. We note that $\sigma = 3$ produces results which are almost indistinguishable from the Dolph-Tschebyscheff solution while the choice $\sigma = 5$ seems to be too large for small spacings.

8.3.2 Two Multi-criteria Versions

In contrast to this formulation as a constrained minimization problem, we consider here the **multi-criteria version** of the Weak Dolph Problem. Following Dolph's comments quoted above, we recognize three conflicting performance criteria: It is desirable to synthesize a pattern with minimal side-lobes and, at the same time, a maximal height for the main lobe and minimal beam-width.

[3] This does not mean that the optimal solutions coincide. However, it assures that if the value of $(\mathcal{P}_{DT})$ is small then so is the value of $(\mathcal{P}_W)$ and vice versa.

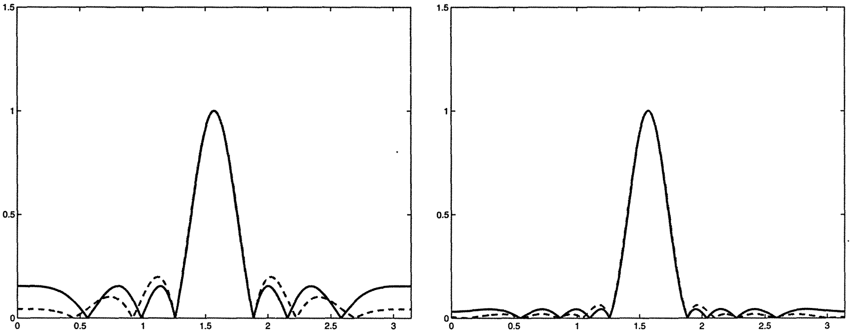


Fig. 8.3. Optimal Array Patterns for $\sigma = 0$

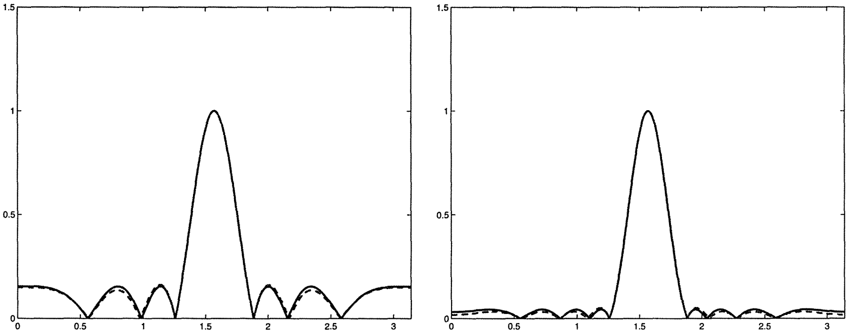


Fig. 8.4. Optimal Array Patterns for $\sigma = 3$

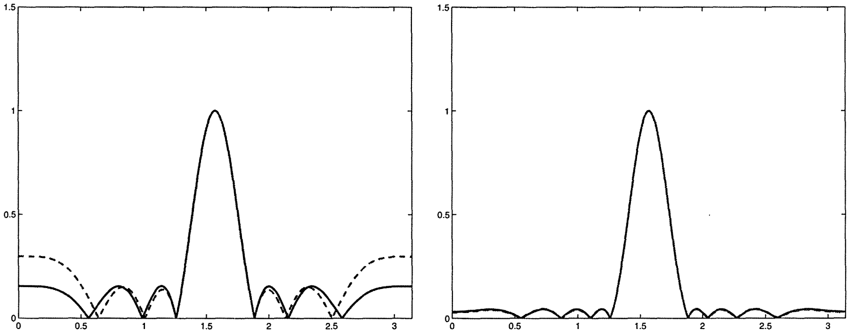


Fig. 8.5. Optimal Array Patterns for $\sigma = 5$

Yet we know from our earlier analysis that maximization of the main beam power also *increases* the power in the side-lobes.

First, we consider the requirements to simultaneously minimize the side-lobes and maximize the height of the main lobe. We define *two* performance indices, the first by

$$F_1(\boldsymbol{a}) \;:=\; -\,\alpha_1\,\left|f(\pi/2)\right|^2 \;=\; -\,\alpha_1\left(\boldsymbol{a}^\top\boldsymbol{b}\right)^2, \tag{8.21a}$$

where, as in (8.18), $b_n = w_n(\pi/2) = 1 - \cos(nkd\cos\hat{\theta})$, and the second by

$$F_2(\boldsymbol{a}) \;:=\; \alpha_2 \int\limits_0^{\hat{\theta}} \mathrm{e}^{-\sigma(\hat{\theta}-\theta)}\,\left|f(\theta)\right|^2\,d\theta \;=\; \alpha_2\,\boldsymbol{a}^\top\boldsymbol{W}\,\boldsymbol{a}, \tag{8.21b}$$

the matrix elements W_{nm} again being given by (8.17). The *vector criterion* is then

$$\boldsymbol{\mathcal{F}}(\boldsymbol{a}) = \left(F_1(\boldsymbol{a})\,,\;F_2(\boldsymbol{a})\,\right)^\top \;\in \mathbb{R}^2, \quad \boldsymbol{a} \in \mathcal{A} \subset \mathbb{R}^n, \tag{8.21c}$$

where $\mathcal{A}$ is some suitable admissible set of feeding coefficients. Note that F_1 was used in describing the constraint in the original formulation of the problem. The positive weights α_1 and α_2 may be chosen in various ways. For the numerical results we give in Example 8.23 below, we will use them to normalize the calculations relative to the uniform feeding. This normalization will allow us to make appropriate comparisons of different methods in a meaningful way.

In this formulation, we are asked to find the set of *Pareto points* for the criterion as explained above. We now apply the results of Subsection 8.2 to the multi-criteria Dolph problem (8.21). If we consider the unconstrained case, i.e., $\mathcal{A} = \mathbb{R}^n$, then there is a unique (up to a multiple constant) Pareto point which coincides with the earlier result (8.19). This is the content of the following theorem.

Theorem 8.21. *Let $\mathcal{A} = \mathbb{R}^n$ and $\boldsymbol{\mathcal{F}} : \mathbb{R}^n \to \mathbb{R}^2$ be given by (8.21a), (8.21b). Then $\boldsymbol{a}^o \neq 0$ is Pareto minimal if and only if*

$$\boldsymbol{a}^o \;=\; \eta\,\boldsymbol{W}^{-1}\boldsymbol{b} \quad \textit{for some } \eta \neq 0\,.$$

Proof: Let $\boldsymbol{a}^o \neq \mathbf{0}$ be a Pareto point and note first that $\boldsymbol{b}^\top\boldsymbol{a}^o \neq 0$ for, if the contrary were true, then $\boldsymbol{b}^\top\boldsymbol{a}^o = 0$ so that $F_1(\boldsymbol{a}^o) = -\,\alpha_1\left({\boldsymbol{a}^o}^\top\boldsymbol{b}\right)^2 = 0$ and we conclude that $F_1(0) = F_1(\boldsymbol{a}^o)$ and $F_2(0) = 0 < F_2(\boldsymbol{a}^o)$ so that $\boldsymbol{a}^o$ cannot be a Pareto point.

Since $\boldsymbol{b}^\top\boldsymbol{a}^o \neq 0$, it follows that

$$\hat{\boldsymbol{a}} \;:=\; \left[\boldsymbol{b}^\top\boldsymbol{a}^o\right]^{-1}\boldsymbol{a}^o \tag{8.22}$$

is optimal for the Weak Dolph Problem $(\mathcal{P}_W)$. Indeed, let $\boldsymbol{a} \in \mathbb{R}^n$ satisfy $\boldsymbol{a}^\top\boldsymbol{b} = 1$. Then

$$F_1\big([\boldsymbol{b}^\top\boldsymbol{a}^o]\,\boldsymbol{a}\big) \;=\; -\,\alpha_1\left(\boldsymbol{b}^\top\boldsymbol{a}^o\right)^2 \;=\; F_1(\boldsymbol{a}^o)\,.$$

Since $\boldsymbol{a}^o$ is a Pareto point by assumption, it follows that

$$F_2\big([\boldsymbol{b}^\top\boldsymbol{a}^o]\,\boldsymbol{a}\big) \;\geq\; F_2(\boldsymbol{a}^o)\,,$$

i.e.,

$$\left(\boldsymbol{b}^\top \boldsymbol{a}^o\right)^2 \boldsymbol{a}^\top \boldsymbol{W} \boldsymbol{a} \;\geq\; \boldsymbol{a}^{o\top} \boldsymbol{W} \boldsymbol{a}^o ,$$

or, using (8.22), $\boldsymbol{a}^\top \boldsymbol{W} \boldsymbol{a} \geq \hat{\boldsymbol{a}}^\top \boldsymbol{W} \hat{\boldsymbol{a}}$. From this inequality and the equation $F_1(\boldsymbol{a}) = F_1(\hat{\boldsymbol{a}})$, the optimality of $\hat{\boldsymbol{a}}$ for $(\mathcal{P}_W)$ follows.

The uniqueness of the optimal solution for $(\mathcal{P}_W)$ then implies that

$$\hat{\boldsymbol{a}} \;=\; [\boldsymbol{b}^\top \boldsymbol{W}^{-1} \boldsymbol{b}]^{-1} \boldsymbol{W}^{-1} \boldsymbol{b} \tag{8.23}$$

i.e.,

$$\boldsymbol{a}^o \;=\; \eta \boldsymbol{W}^{-1} \boldsymbol{b} \quad \text{with} \quad \eta \;=\; \left(\boldsymbol{b}^\top \boldsymbol{a}^o\right) \left[\boldsymbol{b}^\top \boldsymbol{W}^{-1} \boldsymbol{b}\right]^{-1} \;\neq\; 0 .$$

Conversely, let $\boldsymbol{a}^o := \eta \boldsymbol{W}^{-1} \boldsymbol{b}$ for some $\eta \neq 0$. First, we observe that $\boldsymbol{b}^\top \boldsymbol{a}^o = \eta \boldsymbol{b}^\top \boldsymbol{W}^{-1} \boldsymbol{b} \neq 0$ and thus, by (8.19),

$$\hat{\boldsymbol{a}} \;:=\; \frac{1}{\boldsymbol{b}^\top \boldsymbol{a}^o} \boldsymbol{a}^o \;=\; \frac{1}{\boldsymbol{b}^\top \boldsymbol{W}^{-1} \boldsymbol{b}} \boldsymbol{W}^{-1} \boldsymbol{b}$$

is the (unique) optimal solution of $(\mathcal{P}_W)$.

Now let $\boldsymbol{a} \in \mathbb{R}^n$ with

$$\boldsymbol{\mathcal{F}}(\boldsymbol{a}) \;\leq\; \boldsymbol{\mathcal{F}}(\boldsymbol{a}^o) . \tag{8.24}$$

We need to show that this inequality implies $\boldsymbol{\mathcal{F}}(\boldsymbol{a}) = \boldsymbol{\mathcal{F}}(\boldsymbol{a}^o)$. From

$$-\alpha_1 \left(\boldsymbol{b}^\top \boldsymbol{a}\right)^2 \;=\; F_1(\boldsymbol{a}) \;\leq\; F_1(\boldsymbol{a}^o) \;=\; -\alpha_1 \left(\boldsymbol{b}^\top \boldsymbol{a}^o\right)^2 \;<\; 0$$

we conclude that $\boldsymbol{b}^\top \boldsymbol{a} \neq 0$. Therefore,

$$\tilde{\boldsymbol{a}} \;:=\; \frac{1}{\boldsymbol{b}^\top \boldsymbol{a}} \boldsymbol{a}$$

is admissible for $(\mathcal{P}_W)$ and thus

$$\frac{1}{\left(\boldsymbol{b}^\top \boldsymbol{a}\right)^2} \boldsymbol{a}^\top \boldsymbol{W} \boldsymbol{a} \;=\; \tilde{\boldsymbol{a}}^\top \boldsymbol{W} \tilde{\boldsymbol{a}} \;\geq\; \hat{\boldsymbol{a}}^\top \boldsymbol{W} \hat{\boldsymbol{a}} \;=\; \frac{1}{\left(\boldsymbol{b}^\top \boldsymbol{a}^o\right)^2} \boldsymbol{a}^{o\top} \boldsymbol{W} \boldsymbol{a}^o ,$$

i.e.

$$\left(\frac{\boldsymbol{b}^\top \boldsymbol{a}^o}{\boldsymbol{b}^\top \boldsymbol{a}}\right)^2 \boldsymbol{a}^\top \boldsymbol{W} \boldsymbol{a} \;\geq\; \boldsymbol{a}^{o\top} \boldsymbol{W} \boldsymbol{a}^o ,$$

or, since $\alpha_1 \left(\boldsymbol{b}^\top \boldsymbol{a}^o\right)^2 = -F_1(\boldsymbol{a}^o) \leq -F_1(\boldsymbol{a}) = \alpha_1 \left(\boldsymbol{b}^\top \boldsymbol{a}\right)^2$, that

$$F_2(\boldsymbol{a}) \;\geq\; \left(\frac{\boldsymbol{b}^\top \boldsymbol{a}^o}{\boldsymbol{b}^\top \boldsymbol{a}}\right)^2 F_2(\boldsymbol{a}^o) \;\geq\; F_2(\boldsymbol{a}^o) . \tag{8.25}$$

This shows that $F_2(\boldsymbol{a}) \geq F_2(\boldsymbol{a}^o)$ which, together with the assumption (8.24), implies $F_2(\boldsymbol{a}) = F_2(\boldsymbol{a}^o)$.

To show that $F_1(\boldsymbol{a}^o) = F_1(\boldsymbol{a})$ we observe that now equality holds in (8.25). From this and the positive definiteness of $\boldsymbol{W}$ we conclude that $\boldsymbol{b}^\top \boldsymbol{a} = \boldsymbol{b}^\top \boldsymbol{a}^o$ since $\boldsymbol{a}^{o\top} \boldsymbol{W} \boldsymbol{a}^o > 0$. This ends the proof for the unconstrained case. □

Turning now to the constrained problem, let $\mathcal{A}$ be given by

$$\mathcal{A} \;:=\; \big\{\boldsymbol{a} \in \mathbb{R}^n : \|\boldsymbol{a}\|_2 \leq c\big\} \quad \text{for some } c > 0 \tag{8.26}$$

where again $\|\cdot\|_2$ denotes the Euclidean norm in $\mathbb{R}^n$. Existence of Pareto points is assured by Theorem 8.10 since $\mathcal{A}$ is clearly compact and therefore so is $\mathcal{F}(\mathcal{A}) \subset \mathbb{R}^2$.

We now wish to apply the following multiplier rule which is a direct corollary of Theorem 8.16.

Theorem 8.22. *Let $\mathcal{A}$ be defined by (8.26) and $d < \lambda$. Then the functional $\mathcal{F}$ given by (8.21) admits Pareto points, and for every such Pareto point $\boldsymbol{a}^o$ there exist multipliers $\eta_1, \eta_2, \rho \geq 0$ such that*

(i) $\eta_1 + \eta_2 > 0$,
(ii) $\rho\big(\|\boldsymbol{a}^o\| - c\big) = 0$, *and*
(iii) $-\eta_1\big(\boldsymbol{b}^\top \boldsymbol{a}^o\big)\boldsymbol{b} \;+\; \eta_2 \boldsymbol{W}\boldsymbol{a}^o \;+\; \rho\,\boldsymbol{a}^o \;=\; 0\,.$

We now use conditions (i) – (iii) to identify the set of Pareto points for the constrained problem. First, we note that with $\boldsymbol{a}^o$ also $-\boldsymbol{a}^o$ is optimal. We restrict ourselves to the consideration of $\boldsymbol{a}^o$ with $\boldsymbol{b}^\top \boldsymbol{a}^o \geq 0$ and $\boldsymbol{a}^o \neq 0$. Also, we note that $\boldsymbol{b}^\top \boldsymbol{a}^o \neq 0$ since otherwise, for $t \neq 0$, $|t| < 1$,

$$F_1(t\boldsymbol{a}^o) \;=\; 0 \;=\; F_1(\boldsymbol{a}^o)\,, \quad F_2(t\boldsymbol{a}^o) \;=\; t^2\,F_2(\boldsymbol{a}^o) \;<\; F_2(\boldsymbol{a}^o)\,,$$

and $\|t\boldsymbol{a}^o\| \leq \|\boldsymbol{a}^o\| \leq c$ which contradicts the optimality of $\boldsymbol{a}^o$. Therefore, we can replace $\eta_j \geq 0$ by $\eta_j/\big(\boldsymbol{b}^\top \boldsymbol{a}^o\big) \geq 0$, $j = 1, 2, 3$, and have that (i) and (ii) hold and

$$-\eta_1\,\boldsymbol{b} \;+\; \eta_2\,\boldsymbol{W}\boldsymbol{a}^o \;+\; \rho\,\boldsymbol{a}^o \;=\; 0\,. \tag{8.27}$$

We discuss three cases.

Case 1: $\rho = 0$. From (8.27) it follows that

$$\boldsymbol{W}\boldsymbol{a}^o \;=\; \frac{\eta_1}{\eta_2}\,\boldsymbol{b}$$

since $\eta_2 > 0$ (otherwise $\boldsymbol{b} = \boldsymbol{0}$ by (8.27) and (i)). Therefore the line segment

$$\mathcal{S}_1 \;:=\; \left\{\tau\,\boldsymbol{W}^{-1}\boldsymbol{b} : 0 < \tau \leq \frac{c}{\|\boldsymbol{W}^{-1}\boldsymbol{b}\|}\right\} \tag{8.28}$$

is a subset of the set of all critical points.

Case 2: $\rho > 0$ and $\eta_2 = 0$. Then $\|\boldsymbol{a}^o\| = c$ and $\boldsymbol{a}^o = \eta_1/\rho\,\boldsymbol{b}$ by (8.27). Hence

$$a^o = \frac{c}{\|b\|}\, b \tag{8.29}$$

is the only candidate for a Pareto point in this case.

<u>Case 3:</u> $\rho > 0$ and $\eta_2 > 0$. Then again $\|a^o\| = c$ and $\left(W + \tau_3/\eta_2\, I\right) a^o = \eta_1/\eta_2\, b$ by (8.27). Let $\tau = \rho/\eta_2$ and $\hat{a}$ be the solution of

$$(W + \tau I)\,\hat{a} = b\,.$$

(note that a unique solution exists since W, and hence $W + \tau I$, is positive definite.) Then the vector

$$a^o = \frac{c}{\|\hat{a}\|}\, \hat{a}$$

is optimal and so the set $\mathcal{S}_2$ given by

$$\mathcal{S}_2 := \left\{ \frac{c}{\|(W + \tau I)^{-1} b\|}\, (W + \tau I)^{-1} b : \tau \geq 0 \right\} \tag{8.30}$$

is also a subset of the set of critical points.

Combining these sets, we conclude that the set

$$\mathcal{S} = \mathcal{S}_1 \cup \mathcal{S}_2 \cup \left\{ \frac{c}{\|b\|}\, b \right\} \tag{8.31}$$

contains all Pareto points with $b^\top a^o \geq 0$. Finally, we observe that

$$\lim_{\tau \to 0+} \left(\frac{c}{\|(W + \tau I)^{-1} b\|}\, (W + \tau I)^{-1} b \right) = \frac{c}{\|W^{-1} b\|}\, W^{-1} b$$

and

$$\lim_{\tau \to \infty} \left(\frac{c}{\|(W + \tau I)^{-1} b\|}\, (W + \tau I)^{-1} b \right) = \frac{c}{\|b\|}\, b,$$

and hence the set of Pareto points is a subset of a connected one parameter family of points of $\mathbb{R}^n$, i.e. a curve $a^o = a^o(\tau)$ in $\mathbb{R}^n$.

We illustrate the results of the preceeding analysis with several computations whose results are given in the following graphs. For these results, we use specific weights α_1 and α_2 in the definition of the vector cost functional (8.21a) and (8.21b). Specifically, we take

$$F_1(a) := -\frac{f(\pi/2)}{|f_u(\pi/2)|} = -\frac{a^\top b}{|a_u^\top b|}, \tag{8.32a}$$

and

$$F_2(a) := \frac{a^\top W\, a}{a_u^\top W\, a_u}, \tag{8.32b}$$

where $a_u := (1, 1, \ldots, 1)^\top / \sqrt{n}$. Otherwise said, we normalize the vector criteria using the uniform feeding as the reference point.

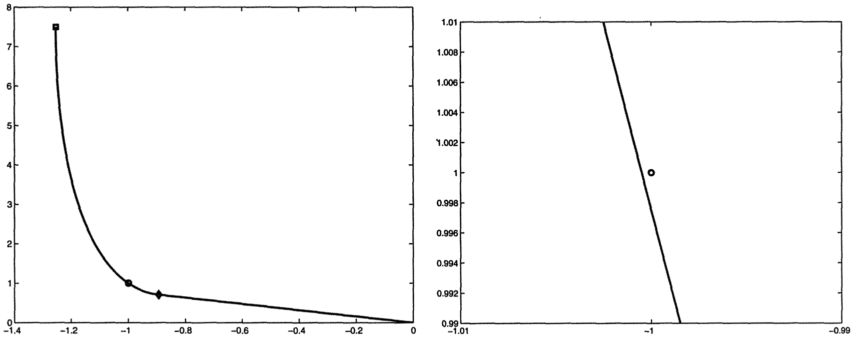


Fig. 8.6. Set of Pareto values of Example 8.23 (3 elements)

Example 8.23. As a first example we consider the configuration of the Weak Dolph Problem from Example 8.20 again, i.e. $n = 3$, $d = \lambda/2$, and $\hat{\theta} = \frac{4}{5} \cdot \frac{\pi}{2}$. As the weight factor we choose $\sigma = 3$, motivated by Example 8.20.

For this configuration (and $c = 1$) we computed now the Pareto points $\boldsymbol{a} = \boldsymbol{a}(\tau)$ of $\mathcal{S}$ and the corresponding values $\mathcal{F}\big(\boldsymbol{a}(\tau)\big)$. The curve $\tau \mapsto \mathcal{F}\big(\boldsymbol{a}(\tau)\big)$ in $\mathbb{R}^2$ is given in Figure 8.6 (left) and consists of two parts. They correspond to $\mathcal{S}_1$ and $\mathcal{S}_2$: The diamond corresponds to parameter $\tau = 0$ in $\mathcal{S}_2$, i.e. to the feeding $\boldsymbol{a} = \big\|\boldsymbol{W}^{-1}\boldsymbol{b}\big\|^{-1}\boldsymbol{W}^{-1}\boldsymbol{b}$. The straight line joining the diamond with the origin are the function values corresponding to $\mathcal{S}_1$. At the other end of the curve, the square marks the value corresponding to $\tau = \infty$ in $\mathcal{S}_2$, i.e. to the feeding $\boldsymbol{a} = \big\|\boldsymbol{b}\big\|^{-1}\boldsymbol{b}$. In the right part of Figure 8.6, part of the same graph is shown in a different scale. The circle marks the point $(-1, 1)$ which corresponds to the uniform feeding. We note that, for this example, the uniform feeding produces an array which is "almost" Pareto optimal.

Figure 8.7 shows a second example for the same spacing d and beam-width but $n = 5$ elements. Here we see clearly, that the uniform feeding is not even close to optimal.

A *second possibility* to treat the Dolph problem as a vector optimization problem is to simultaneously minimize the beam-width $2(\pi/2 - \hat{\theta})$ and the side-lobes while requiring that the height of the main lobe be normalized to one. In our formulation, $\hat{\theta}$ and the feeding coefficients a_n are the unknowns, and the vector optimization problem is formulated as

$$(\mathcal{P}_{vec}) \qquad \begin{array}{ll} \text{Minimize} & \mathcal{F}(\hat{\theta}, \boldsymbol{a}) \\ \text{subject to} & \hat{\theta} \in [0, \pi/2]\,, \ \boldsymbol{a} \in \mathbb{R}^n\,, \ \boldsymbol{b}(\hat{\theta})^\top \boldsymbol{a} = 1 \end{array}$$

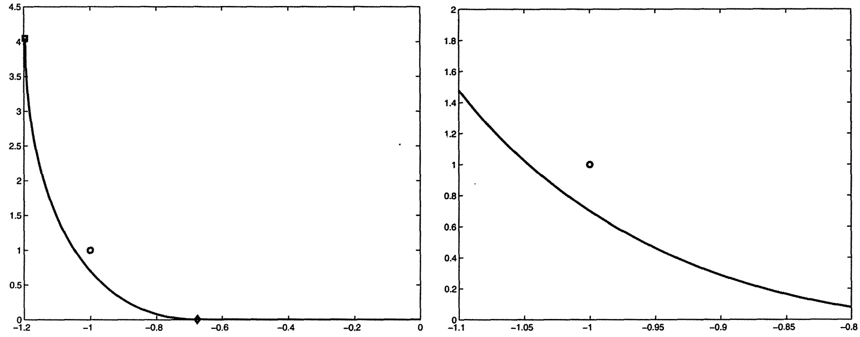


Fig. 8.7. Set of Pareto values of Example 8.23 (5 elements)

where

$$\mathcal{F}(\hat{\theta}, \boldsymbol{a}) \;=\; \big(F_1(\hat{\theta}, \boldsymbol{a}), F_2(\hat{\theta}, \boldsymbol{a})\big)^\top$$

and

$$F_1(\hat{\theta}, \boldsymbol{a}) \;=\; \frac{\pi}{2} - \hat{\theta}, \quad F_2(\hat{\theta}, \boldsymbol{a}) \;=\; \boldsymbol{a}^\top \boldsymbol{W}(\hat{\theta})\boldsymbol{a}. \tag{8.33}$$

The vector-valued and matrix-valued, respectively, functions $\boldsymbol{b}(\hat{\theta})$ and $\boldsymbol{W}(\hat{\theta})$ are defined by

$$b_m(\hat{\theta}) \;=\; 1 \;-\; \cos(mkd\cos\hat{\theta}), \quad m = 1, \dots, n,$$

and, for $\ell, m = 1, \dots, n$,

$$W(\hat{\theta})_{\ell m} \;=\; \int_0^{\hat{\theta}} \mathrm{e}^{-\sigma(\hat{\theta}-\theta)} \left[\cos(\ell kd\cos\theta) - \cos(\ell kd\cos\hat{\theta})\right] \cdot$$

$$\cdot \left[\cos(mkd\cos\theta) - \cos(mkd\cos\hat{\theta})\right] d\theta\,.$$

Obviously, $(\hat{\theta}^o, \boldsymbol{a}^o) = (0, \boldsymbol{a}^o)$ with any $\boldsymbol{a}^o \in \mathbb{R}^n$ such that $\boldsymbol{b}(0)^\top \boldsymbol{a}^o = 1$ is Pareto optimal since $F_2(0, \boldsymbol{a}^o) = 0$. On the other hand, there exists no $\boldsymbol{a}$ such that $(\pi/2, \boldsymbol{a})$ is even admissible since $\boldsymbol{b}(\pi/2) = \boldsymbol{0}$. In addition to the extreme situation $\hat{\theta} = 0$ there exist Pareto points:

Theorem 8.24. *There exist Pareto points of $(\mathcal{P}_{vec})$ with $\hat{\theta} > 0$. If $d < \lambda$ then all of the Pareto points $\boldsymbol{a}^o$ with $\hat{\theta}^o \in (0, \pi/2)$ necessary have the form*

$$\boldsymbol{a}^o \;=\; \left[\boldsymbol{b}(\hat{\theta})^\top \boldsymbol{W}(\hat{\theta})^{-1} \boldsymbol{b}(\hat{\theta})\right]^{-1} \boldsymbol{W}(\hat{\theta})^{-1} \boldsymbol{b}(\hat{\theta})\,. \tag{8.34}$$

The values of $\mathcal{F}$ lie on the curve parametrized by $\hat{\theta}$

$$\hat{\theta} \;\mapsto\; \left(\frac{\pi}{2} - \hat{\theta}\,,\; \frac{1}{\boldsymbol{b}(\hat{\theta})^\top \boldsymbol{W}(\hat{\theta})^{-1} \boldsymbol{b}(\hat{\theta})}\right)^\top \in \mathbb{R}^2\,, \quad \hat{\theta} \in (0, \pi/2)\,.$$

Proof: First, we note that $\boldsymbol{W}(\theta)$ and $\boldsymbol{b}(\theta)$ depend continuously on θ on the interval $[0, \pi/2]$. Since $\boldsymbol{W}(\theta)$ is positive definite for every $\theta \in (0, \pi/2]$ we conclude that $\boldsymbol{W}(\theta)$ is uniformly positive definite on $[\varepsilon, \pi/2]$ for every $\varepsilon > 0$, i.e. there exists $\gamma = \gamma(\varepsilon) > 0$ with $\boldsymbol{a}^\top \boldsymbol{W}(\theta)\boldsymbol{a} \geq \gamma \|\boldsymbol{a}\|^2$ for all $\boldsymbol{a} \in \mathbb{R}^n$ and all $\theta \in [\varepsilon, \pi/2]$. To show existence of Pareto points we apply Theorem 8.10 and so have to prove compactness of the set

$$\mathcal{T} := \left\{\boldsymbol{\mathcal{F}}(\hat{\theta}, \boldsymbol{a}) \in \mathbb{R}^2 : 0 \leq \hat{\theta} \leq \pi/2,\ \boldsymbol{b}(\hat{\theta})^\top \boldsymbol{a} = 1,\ F_1(\hat{\theta}, \boldsymbol{a}) \leq z_1,\ F_2(\hat{\theta}, \boldsymbol{a}) \leq z_2\right\}$$

for some $(z_1, z_2) \in \mathbb{R}^2$. We choose $z_1 \in (0, \pi/2)$ and $z_2 > 0$. The set $\mathcal{T}$ is certainly bounded since $0 \leq F_1(\hat{\theta}, \boldsymbol{a}) \leq z_1$ and $0 \leq F_2(\hat{\theta}, \boldsymbol{a}) \leq z_2$ for all $\boldsymbol{\mathcal{F}}(\hat{\theta}, \boldsymbol{a}) \in \mathcal{T}$. To verify that $\mathcal{T}$ is closed let $\boldsymbol{\mathcal{F}}(\hat{\theta}_j, \boldsymbol{a}_j) \to \boldsymbol{u}$, $j \to \infty$. Then $u_1 \leq z_1 < \pi/2$, i.e. $F_1(\hat{\theta}_j, \boldsymbol{a}_j) \to u_1$ implies $\hat{\theta}_j \to \hat{\theta} := \pi/2 - u_1 > 0$. The second component yields $\boldsymbol{a}_j^\top \boldsymbol{W}(\hat{\theta}_j)\boldsymbol{a}_j \to u_2$, $j \to \infty$. We have $\hat{\theta}_j \geq \varepsilon := \hat{\theta}/2 > 0$ for sufficiently large j. From the estimate $\gamma(\varepsilon) \|\boldsymbol{a}_j\|^2 \leq \boldsymbol{a}_j^\top \boldsymbol{W}(\hat{\theta}_j)\boldsymbol{a}_j$ we conclude that the sequence $\{\boldsymbol{a}_j\}$ is bounded and contains a convergent subsequence, again denoted by $\{\boldsymbol{a}_j\}$, i.e. $\boldsymbol{a}_j \to \boldsymbol{a}$ for some $\boldsymbol{a} \in \mathbb{R}^n$. Finally, the continuity of $\boldsymbol{W}$ and $\boldsymbol{b}$ yields $\boldsymbol{b}(\hat{\theta})^\top \boldsymbol{a} = 1$ and $\boldsymbol{a}^\top \boldsymbol{W}(\hat{\theta})\boldsymbol{a} = u_2$. This, together with the definition of $\hat{\theta}$, yields the closedness of $\mathcal{T}$. Therefore, $\boldsymbol{\mathcal{F}}$ has Pareto points $(\hat{\theta}^o, \boldsymbol{a}^o)$ with $\hat{\theta}^o > 0$ since $F_1(\hat{\theta}^o, \boldsymbol{a}^o) \leq z_1 < \pi/2$.

Now let $\left(\hat{\theta}^o, \boldsymbol{a}^o\right)$ be a Pareto point with $\hat{\theta}^o \neq 0, \pi/2$. We note that $(\mathcal{P}_{vec})$ has the form in which the Lagrange multiplier rule of Theorem 8.16 can be applied. This yields existence of $\eta_1, \eta_2, \rho \in \mathbb{R}$ with $\eta_1 + \eta_2 > 0$ and

$$-\eta_1 \ + \ \eta_2\, \boldsymbol{a}^{o\top} \boldsymbol{W}'(\hat{\theta}^o)\boldsymbol{a}^o \ + \ \rho\, \boldsymbol{b}'(\hat{\theta}^o)^\top \boldsymbol{a}^o = 0\,, \tag{8.35a}$$

$$2\,\eta_2\, \boldsymbol{W}(\hat{\theta}^o)\boldsymbol{a}^o \ + \ \rho\, \boldsymbol{b}(\hat{\theta}^o) = 0\,. \tag{8.35b}$$

We note that $\eta_2 > 0$ since otherwise also $\rho \neq 0$ and thus $\boldsymbol{b}(\hat{\theta}^o) = 0$ which can only happen for the $\hat{\theta}^o = \pi/2$ which we excluded. Indeed, from $b_1(\hat{\theta}^o) = 0$ we note that $kd \cos \hat{\theta}^o = 2\pi m$ for some m, i.e. $\frac{d}{\lambda} \cos \theta^o = m$. From $d/\lambda < 1$ this is only possible for $\cos \theta^o = 0$, i.e. $\hat{\theta}^o = \pi/2$. Therefore, $\boldsymbol{a}^o$ has the form $\boldsymbol{a}^o = \tau\, \boldsymbol{W}(\hat{\theta}^o)^{-1} \boldsymbol{b}(\hat{\theta}^o)$ with $\tau = -\rho/(2\eta_2) \in \mathbb{R}$. We determine τ by the normalization $\boldsymbol{b}(\hat{\theta}^o)^\top \boldsymbol{a}^o = 1$ which yields the form (8.34). This ends the proof. □

We remark that (8.34) has the same form as the solution of the Weak Dolph Problem $(\mathcal{P}_W)$ for fixed $\hat{\theta}^o \in (0, \pi/2)$. This result confirms that, in the multi-criteria setting, just as observed in Dolph's original paper, there are equivalent formulations of the optimization problem.

From (8.34) we can easily show that

$$\lim_{\hat{\theta} \to 0} \frac{1}{\boldsymbol{b}(\hat{\theta})^\top \boldsymbol{W}(\hat{\theta})^{-1} \boldsymbol{b}(\hat{\theta})} = 0 \quad \text{and}$$

$$\lim_{\hat{\theta} \to \pi/2} \frac{1}{\boldsymbol{b}(\hat{\theta})^\top \boldsymbol{W}(\hat{\theta})^{-1} \boldsymbol{b}(\hat{\theta})} = \infty\,.$$

Indeed, defining $\boldsymbol{a} = \|\boldsymbol{b}(\hat{\theta})\|_2^{-2}\,\boldsymbol{b}(\hat{\theta})$ we note that $\boldsymbol{a}^\top \boldsymbol{b}(\hat{\theta}) = 1$ and thus from the optimality of $\boldsymbol{a}^o$ and (8.20)

$$\frac{1}{\boldsymbol{b}(\hat{\theta})^\top \boldsymbol{W}(\hat{\theta})^{-1}\boldsymbol{b}(\hat{\theta})} = F_2(\boldsymbol{a}^o) \leq F_2(\boldsymbol{a}) = \frac{1}{\|\boldsymbol{b}(\hat{\theta})\|_2^4\,\boldsymbol{b}(\hat{\theta})^\top \boldsymbol{W}(\hat{\theta})\boldsymbol{b}(\hat{\theta})}$$
$$\leq \frac{1}{\|\boldsymbol{b}(\hat{\theta})\|_2^2}\,\|\boldsymbol{W}(\hat{\theta})\|$$

which tends to zero as $\hat{\theta} \to 0$ since $\boldsymbol{W}(0) = 0$. The second limit is seen by the observation that

$$\lim_{\hat{\theta}\to\pi/2}\left[\boldsymbol{b}(\hat{\theta})^\top \boldsymbol{W}(\hat{\theta})^{-1}\boldsymbol{b}(\hat{\theta})\right] = 0$$

since $\boldsymbol{b}(\pi/2) = \boldsymbol{0}$ and $\boldsymbol{W}(\pi/2)$ is regular.

Let us conclude by recapping the discussion of this section. In contrast with the standard approach of Dolph in which feeding coefficients are sought to minimize the peak side-lobe power under the constraint of fixed beam width and peak main beam power, we consider *both* the side-lobe power and the main beam power or the side-lobe power and the beam width as quantities to be optimized. The mathematical theory provides conditions for computing the feeding coefficients which are *Pareto optimal.* The set of Pareto points in this case then gives a tradeoff curve for the Dolph array which shows exactly what price is paid in main beam power level reduction for a desired decrease in side-lobe power. Such curves or surfaces in higher dimensions, should be valuable information for the antenna designer. The mathematical tools are available to treat objective functionals of higher dimension as well. In the next sections, we apply these ideas to more complicated problems and illustrate the method with numerical results.

8.4 Null Placement Problems and Super-gain

In Section 7.2 we considered several problems which relate to the need to feed an antenna so that the radiated power is maximized either over a continuous sector, or in one or more specific directions. We can write this constrained optimization problem as (see (7.2))

$$\text{Maximize} \quad \|\alpha\mathcal{K}\psi\|^2_{L^2(S^{d-1})} \quad \text{subject to} \quad \psi \in U\,. \tag{8.36}$$

The set U represents constraints and we have considered, among others, the simple norm constraint $\|\psi\|_X \leq 1$.

One might hope that by maximizing the power in a given sector of the far field by choosing some optimal feeding $\psi^o \in U$, the power over the *complementary* portion of S^{d-1} is thereby minimized. The numerical results in the work of

Fast [40] show that this hope is frustrated. Considering the examples of array patterns in Section 1.2, this should come as no surprise.

While the Dolph problem, as we have treated it, involves both an integral functional and a point-evaluation functional, we turn now to vector criteria whose components involve various integral functionals which have been individually discussed in Chapters 3, 4, and 7. Some of the constrained optimization problems studied earlier can be studied now as multi-criteria problems. Four of the possibilities are listed in the table below using our familiar notation.

Problem	Description	Vector Criterion
I	Minimal pattern deviation with null placement	$\mathcal{F} = \left(\|\alpha[\mathcal{K}\psi - f^o]\|^2\,,\ \|\beta\mathcal{K}\psi\|^2\right)^\top$
II	Maximum directed power with null placement	$\mathcal{F} = \left(-\|\alpha\mathcal{K}\psi\|^2\,,\ \|\beta\mathcal{K}\psi\|^2\right)^\top$
III	Maximum directed power with minimal super-gain	$\mathcal{F} = \left(-\|\alpha\mathcal{K}\psi\|^2\,,\ \|\psi\|^2/\|\mathcal{K}\psi\|^2\right)^\top$
IV	Maximum directed power with null placement and minimal super-gain	$\mathcal{F} = \left(-\|\alpha\mathcal{K}\psi\|^2\,,\ \|\beta\mathcal{K}\psi\|^2\,,\ \frac{\|\psi\|^2}{\|\mathcal{K}\psi\|^2}\right)^\top$

For example in Subsection 7.3.1 we considered the problem of maximizing the power in a prescribed sector with constraints on the power in another, non-intersecting, sector i.e., for some prescribed constant $c > 0$, $\|\beta\mathcal{K}\psi\| \leq c$. We suggest that, as in the multi-criteria Dolph problem, since it is difficult to set the constraint level *a priori*, we pose the problem as a constrained multi-criteria problem.

To do so, we select a sector, $\mathcal{A}$, in which we wish to maximize the far field power and a sector (or sectors), $\mathcal{B}$, in which we wish to minimize that power. We can interpret this latter choice as specifying the directions in which we desire to "place nulls". We then write the constrained multi-criteria problem as:

$$\begin{array}{llll} \text{Minimize} & \mathcal{F}(\psi) & = \begin{pmatrix} F_1(\psi) \\ F_2(\psi) \end{pmatrix} := \begin{pmatrix} -\|\alpha\mathcal{K}\psi\|^2_{L^2(S^{d-1})} \\ \|\beta\mathcal{K}\psi\|^2_{L^2(S^{d-1})} \end{pmatrix} \\ \text{subject to} & \|\psi\|_X & \leq 1\,. \end{array} \tag{8.37}$$

The weighting functions α and β specify the different directions in which we want to maximize and minimize power, respectively. We can think of α and β to be the characteristic functions of the sections $\mathcal{A}$ and $\mathcal{B}$, respectively. In what follows we require only that $\alpha, \beta \in L^\infty(S^{d-1})$ are real-valued.[4] Formulated in this way, we see that this problem (8.37) is one form of the *null placement problem* that we studied in Section 7.3.

This, of course, is only one possible problem that we can study. Indeed, since it is closely related to the Dolph problem already treated, we would rather turn to two of the other problems that are listed in the table.

Before doing so we should make clear the underlying assumptions that will be in force throughout this section. As is Chapter 7, we make the assumptions (A1), (A2), and (A5), so that the compact far field operator is one-to-one with dense range, the functions in the range being analytic. Moreover, we assume an extended form of (A3), namely that the supports of the functions α and β each contain an open set, O_α and O_β respectively, with $O_\alpha \cap O_\beta = \emptyset$.

To begin, let us consider the first problem in the table, namely that of placing low power in some prescribed sector of the far field, while preserving a nominal signal as much as possible in another. This is the problem which we initially treated in Subsection 7.3.4 and for which we gave concrete numerical results in Example 7.9 for the special case of the circular line source.

8.4.1 Minimal Pattern Deviation

The multi-criteria version of this optimization problem is that of finding Pareto points for the objective function

$$\mathcal{F}(\psi) \;=\; \left(\|\alpha(\mathcal{K}\psi - f^o)\|^2_{L^2(S^{d-1})}, \|\beta\mathcal{K}\psi\|^2_{L^2(S^{d-1})}\right)^\top .$$

Since the given far field pattern is in the range of $\mathcal{K}$ we have $\mathcal{K}\hat{\psi} = f^o$ and so it is reasonable to take the norm constraint $\|\psi\|_X \leq c$ where $c \leq \|\hat{\psi}\|_X$.

The question of existence of Pareto points is trivial: Obviously, $\psi^o = 0$ and $\psi^o = \hat{\psi}$ are Pareto points since they yield $F_2(\psi^o) = 0$ and $F_1(\psi^o) = 0$, respectively. It is the task to determine *all* Pareto points. We observe that the functionals F_j, $j = 1, 2$, are convex by Lemma 3.32 as is the constraint functional $g(\psi) = \|\psi\|^2_X - c^2$. Therefore, the necessary optimality conditions from Theorem 8.16 are also sufficient by Theorem 8.17. In order to apply the theorems we need to compute the gradients of the component functionals. By Lemma 3.32 the gradients of $F_1(\psi) = \|\alpha(\mathcal{K}\psi - f^o)\|^2_{L^2(S^{d-1})}$ and $F_2(\psi) = \|\beta\mathcal{K}\psi\|^2_{L^2(S^{d-1})}$ are given by

$$\nabla F_1(\psi) \;=\; 2\,\mathcal{K}^*\alpha^2(\mathcal{K}\psi - f^o)\,, \quad \text{and} \quad \nabla F_2(\psi) \;=\; 2\,\mathcal{K}^*\beta^2\mathcal{K}\psi\,. \tag{8.38}$$

Moreover, the constraint function $g(\psi) = \|\psi\|^2_X - c^2$ has gradient $\nabla g(\psi) = 2\psi$.

[4] Recall that, because of the analyticity of the far field, it is not possible to require that the far field actually vanish on a subset of S^{d-1} with an accumulation point. The formulation here lowers the far field power in those directions and, as numerical results demonstrate, lowers the power substantially.

We exclude the obvious Pareto points $\psi^o = 0$ and $\psi^o = \hat{\psi}$ from the subsequent discussion. From Theorem 8.17 all *weak* Pareto points are given by the solutions of the equations

$$\mathcal{K}^*(\eta_1\alpha^2 + \eta_2\beta^2)\mathcal{K}\psi^o \;+\; \rho\,\psi^o \;=\; \eta_1\,\mathcal{K}^*\alpha^2 f^o \;=\; \eta_1\,\mathcal{K}^*\alpha^2\mathcal{K}\hat{\psi} \tag{8.39a}$$

and

$$\rho\big(\|\psi^o\|_X - c\big) \;=\; 0 \tag{8.39b}$$

for some $\eta_1, \eta_2, \rho \geq 0$ with $\eta_1 + \eta_2 > 0$. The *strong* Pareto points correspond to choices $\eta_1 > 0$ and $\eta_2 > 0$. In the discussion of (8.39a), (8.39b) we consider first the case $\eta_1 = 0$. Then (8.39a) reduces to

$$\eta_2\,\mathcal{K}^*\beta^2\mathcal{K}\psi^o \;+\; \rho\,\psi^o \;=\; 0\,.$$

Multiplication of this equation by ψ^o yields

$$\eta_2\,\|\beta\mathcal{K}\psi^o\|^2_{L^2(S^{d-1})} \;+\; \rho\,\|\psi^o\|^2_X \;=\; 0\,,$$

i.e. both terms have to vanish separately. We note that $\eta_2 > 0$, i.e. $\beta\mathcal{K}\psi^o = 0$. Since $\mathcal{K}\psi^o$ is analytic and $\mathcal{K}$ is one-to-one this leads to the case $\psi^o = 0$.

Now let $\eta_1 > 0$. By dividing (8.39a) by $\eta_1 + \eta_2$ we observe that the pair of equations (8.39a), (8.39b) has the form

$$\mathcal{K}^*\big(\eta\,\alpha^2 + (1-\eta)\,\beta^2\big)\mathcal{K}\psi^o \;+\; \rho\,\psi^o \;=\; \eta\,\mathcal{K}^*\alpha^2 f^o \;=\; \eta\,\mathcal{K}^*\alpha^2\mathcal{K}\hat{\psi}\,. \tag{8.40a}$$

$$\rho \;=\; 0 \quad \text{or} \quad \|\psi^o\|_X \;=\; c \tag{8.40b}$$

where $\eta = \eta_1/(\eta_1 + \eta_2)$. We show the following lemma:

Lemma 8.25. *Assume that $\hat{\psi} \neq 0$. For all $\eta \in (0,1]$ there exist unique $\rho > 0$ and $\psi^o \in X$ (depending on η) which solve (8.40a) and $\|\psi^o\|_X = c$.*

Proof: First we recall that the density of the range of $\mathcal{K}$ in $L^2(S^{d-1})$ is equivalent to the injectivity of the adjoint $\mathcal{K}^*$.

We fix $\eta \in (0,1]$ and note that for all $\rho > 0$ the second kind equation (8.40a) has a unique solution ψ_ρ, i.e. ψ_ρ solves

$$\mathcal{K}^*\big(\eta\,\alpha^2 + (1-\eta)\,\beta^2\big)\mathcal{K}\psi_\rho \;+\; \rho\,\psi_\rho \;=\; \eta\,\mathcal{K}^*\alpha^2 f^o \;=\; \eta\,\mathcal{K}^*\alpha^2\mathcal{K}\hat{\psi}\,. \tag{8.41}$$

The function $\eta\,\alpha^2 + (1-\eta)\,\beta^2$ is non-negative. We set $\gamma = \sqrt{\eta\,\alpha^2 + (1-\eta)\,\beta^2}$ for convenience and show that the mapping $\rho \mapsto \|\psi_\rho\|_X$ is continuous and strictly decreasing for $\rho > 0$ with

$$\lim_{\rho\to\infty}\|\psi_\rho\|_X \;=\; 0 \quad \text{and} \quad \lim_{\rho\to 0}\|\psi_\rho\|_X \;=\; \begin{cases} \infty\,, & \text{if } \eta < 1\,, \\ \|\hat{\psi}\|_X\,, & \text{if } \eta = 1\,. \end{cases}$$

An application of the Intermediate Value Theorem will then yield the assertion.

Multiplication of (8.41) with ψ_ρ yields

$$\|\gamma\mathcal{K}\psi_\rho\|^2_{L^2(S^{d-1})} \;+\; \rho\,\|\psi_\rho\|^2_X \;=\; \eta\left(\mathcal{K}^*\alpha^2 f^o, \psi_\rho\right)_X \;\leq\; \eta\,\|\mathcal{K}^*\alpha^2 f^o\|_X\|\psi_\rho\|_X\,,$$

from which it follows that $\rho\|\psi_\rho\|^2_X \leq \eta\,\|\mathcal{K}^*\alpha^2 f^o\|_X\|\psi_\rho\|_X$ and thus

$$\|\psi_\rho\|_X \;\leq\; \frac{\eta\,\|\mathcal{K}^*\alpha^2 f^o\|_X}{\rho} \;\longrightarrow\; 0 \tag{8.42}$$

as ρ tends to infinity.

Now let $\rho_1, \rho_2 > 0$ and subtract the equations (8.41) for (ρ_1, ψ_1) and (ρ_2, ψ_2) where we write $\psi_j = \psi_{\rho_j}$, $j = 1, 2$. This yields

$$\mathcal{K}^*\gamma^2\mathcal{K}(\psi_1 - \psi_2) \;+\; \rho_1(\psi_1 - \psi_2) \;=\; (\rho_2 - \rho_1)\psi_2\,. \tag{8.43}$$

Multiplication by $\psi_1 - \psi_2$ leads to the equation

$$\|\gamma\mathcal{K}(\psi_1 - \psi_2)\|^2_{L^2(S_1)} \;+\; \rho_1\|\psi_1 - \psi_2\|^2_X \;=\; (\rho_2 - \rho_1)\left(\psi_2, \psi_1 - \psi_2\right)_X \tag{8.44}$$

and thus

$$\rho_1\|\psi_1 - \psi_2\|^2_X \;\leq\; |\rho_2 - \rho_1|\left|\left(\psi_2, \psi_1 - \psi_2\right)_X\right| \;\leq\; |\rho_2 - \rho_1|\,\|\psi_2\|_X\,\|\psi_1 - \psi_2\|_X\,,$$

i.e.

$$\rho_1\|\psi_1 - \psi_2\|_X \;\leq\; |\rho_2 - \rho_1|\,\|\psi_2\|_X \;\leq\; |\rho_2 - \rho_1|\,\frac{\eta\,\|\mathcal{K}^*\alpha^2 f^o\|_X}{\rho_2}$$

by (8.42). This inequality shows that $\rho \mapsto \|\psi_\rho\|_X$ is continuous.

Now take $\rho_2 > \rho_1 > 0$. We note that $\psi_2 \neq 0$ by (8.41) and the fact that $\eta\,\mathcal{K}^*\alpha^2\mathcal{K}\hat{\psi} \neq 0$ by the assumption that $\hat{\psi} \neq 0$. Therefore, (8.43) yields $\psi_1 \neq \psi_2$. From (8.44) we conclude that $\left(\psi_2, \psi_1 - \psi_2\right)_X > 0$ i.e. that

$$\|\psi_2\|^2_X \;<\; \left(\psi_2, \psi_1\right)_X \;\leq\; \|\psi_2\|_X\,\|\psi_1\|_X\,.$$

Therefore $\|\psi_2\|_X < \|\psi_1\|_X$, i.e. the map $\rho \mapsto \|\psi_\rho\|_X$ is strictly monotone decreasing.

It remains to compute $\lim_{\rho\to\infty}\|\psi_\rho\|_X$. Let $\{\rho_j\}$ be any sequence in $\mathbb{R}_{>0}$ converging to zero with corresponding sequence $\{\psi_j\}$. Assume that $\lim_{j\to\infty}\|\psi_j\|_X < \infty$. Then there exists a weakly converging subsequence of $\{\psi_j\}$ with $\psi_j \rightharpoonup \psi$ for some ψ. The compactness of $\mathcal{K}$ guarantees that $\mathcal{K}^*\gamma^2\mathcal{K}\psi_j \to \mathcal{K}^*\gamma^2\mathcal{K}\psi$ in $L^2(S^{d-1})$ and thus $\mathcal{K}^*\gamma^2\mathcal{K}\psi = \eta\,\mathcal{K}^*\alpha^2\mathcal{K}\hat{\psi}$ by (8.41), i.e.

$$\mathcal{K}^*\left(\eta\,\alpha^2 + (1-\eta)\,\beta^2\right)\mathcal{K}\psi \;=\; \eta\,\mathcal{K}^*\alpha^2\mathcal{K}\hat{\psi}\,. \tag{8.45}$$

First let $\eta = 1$. Then we multiply (8.45) by $\psi - \hat{\psi}$, arrive at $\|\alpha\mathcal{K}(\psi - \hat{\psi})\|^2 = 0$, and $\psi = \hat{\psi}$ follows by now familiar arguments. We show norm convergence $\psi_j \to \hat{\psi}$ by writing (8.41) for $\rho = 0$ in the form

$$\mathcal{K}^*\alpha^2\mathcal{K}(\psi_j - \hat{\psi}) + \rho_j(\psi_j - \hat{\psi}) = -\rho_j\hat{\psi},$$

i.e.

$$\begin{aligned}\big\|\alpha\mathcal{K}(\psi_j - \hat{\psi})\big\|^2_{L^2(S^{d-1})} + \rho_j\,\big\|\psi_j - \hat{\psi}\big\|^2_X &= -\rho_j\,\big(\hat{\psi}\,,\,\psi_j - \hat{\psi}\big)_X \\ &\le \rho_j\,\big|\big(\hat{\psi}\,,\,\psi_j - \hat{\psi}\big)_X\big|\,,\end{aligned}$$

and thus

$$\big\|\psi_j - \hat{\psi}\big\|^2_X \le \big|\big(\hat{\psi}\,,\,\psi_j - \hat{\psi}\big)_X\big| \longrightarrow 0$$

as j tends to infinity by the weak convergence $\psi_j \rightharpoonup \hat{\psi}$.
Now let $\eta < 1$. Since $\mathcal{K}^*$ is one-to-one we can write (8.45) in the form

$$\eta\,\alpha^2\mathcal{K}(\psi - \hat{\psi}) + (1-\eta)\,\beta^2\mathcal{K}\psi = 0\,.$$

Restricting this equation to some open set $O \subset (\operatorname{supp}\beta \setminus \operatorname{supp}\alpha)$ yields $\beta^2\mathcal{K}\psi = 0$ on O, i.e. $\psi = 0$ by the analyticity of $\mathcal{K}\psi$ and the injectivity of $\mathcal{K}$. This contradicts the monotonicity of $\rho \mapsto \|\psi_\rho\|_X$ and ends the proof of the lemma. □

We note from the previous proof that the assumption that $\mathcal{K}$ has dense range in $L^2(S^{d-1})$ has been needed only to derive (8.45) i.e. to assure that $\|\psi_\rho\|_X \to \infty$ as ρ tends to zero in the case $\eta < 1$.

This lemma shows again that the solution set of (8.39a), (8.39b) forms a one-dimensional manifold parametrized by $\eta \in (0,1]$.

We now turn to the finite dimensional approximation. Once again, let $\{X_n\}$ be a sequence of finite dimensional subspaces such that $\bigcup_n X_n$ is dense in X. We note by Lemma 3.24 that $\bigcup_n(U \cap X_n)$ is dense in U where U is the ball of radius 1 in X in our particular case. The finite dimensional problem is now formulated as:

$$\begin{aligned}&\text{Find Pareto points of}\\ &\mathcal{F}(\psi) = \left(\|\alpha(\mathcal{K}\psi - f^o)\|^2_{L^2(S^{d-1})}, \|\beta\mathcal{K}\psi\|^2_{L^2(S^{d-1})}\right)^\top \\ &\text{Subject to } \psi \in X_n\,,\ \|\psi\|_X \le c\,.\end{aligned} \tag{8.46}$$

Note that we have not approximated either f^o or $\mathcal{K}$. Therefore, this problem has exactly the same form as before with X replaced by the finite dimensional space X_n. The operator $\mathcal{K}|_{X_n} : X_n \to L^2(S^{d-1})$ is still one-to-one, the images $\mathcal{K}\psi$ are still analytic on S^{d-1} but the range of $\mathcal{K}|_{X_n}$ is no longer dense in $L^2(S^{d-1})$. Let $P_n : X \to X_n$ be the orthogonal projection from X onto X_n. It is characterized by $\big(\psi_n, P_n\tilde{\psi}\big)_X = \big(\psi_n, \tilde{\psi}\big)_X$ for all $\tilde{\psi} \in X$ and $\psi_n \in X_n$. From

$$\big(\mathcal{K}\psi_n, \varphi\big)_{L^2(S^{d-1})} = \big(\psi_n, \mathcal{K}^*\varphi\big)_X = \big(\psi_n, P_n\mathcal{K}^*\varphi\big)_X$$

for all $\psi_n \in X_n$ and $\varphi \in L^2(S^{d-1})$ we observe that the adjoint $(\mathcal{K}|_{X_n})^* : L^2(S^{d-1}) \to X_n$ of $\mathcal{K}|_{X_n} : X_n \to L^2(S^{d-1})$ is given by $P_n\mathcal{K}^*$.

Let $\psi_n^o \neq 0$ be a Pareto point of (8.46). In this case the Lagrange multiplier equations (8.40a), (8.40b) take the form

$$P_n\mathcal{K}^*\big(\eta\,\alpha^2 + (1-\eta)\,\beta^2\big)\mathcal{K}\psi_n^o \;+\; \rho\,\psi_n^o \;=\; \eta\,P_n\mathcal{K}^*\alpha^2 f^o\,. \tag{8.47a}$$

$$\rho \;=\; 0 \quad \text{or} \quad \|\psi_n^o\|_X \;=\; c\,. \tag{8.47b}$$

for $\eta \in [0,1]$. Equation (8.47a) is just equation (8.40a) projected onto X_n. Again, $\eta \neq 0$ since otherwise $P_n\mathcal{K}^*\beta^2\mathcal{K}\psi_n^o + \rho\psi_n^o = 0$. Multiplication with ψ_n^o yields $\|\beta\mathcal{K}\psi_n^o\|^2_{L^2(S^{d-1})} + \rho\|\psi_n^o\|_X = 0$ and thus $\psi_n^o = 0$.

The operator $\psi_n \mapsto P_n\mathcal{K}^*\big(\eta\alpha^2 + (1-\eta)\beta^2\big)\mathcal{K}\psi_n$ from X_n into itself is one-to-one as before and therefore onto since X_n is finite dimensional. This is the basic difference from the infinite dimensional case! Again, if $P_n\mathcal{K}^*\alpha^2\mathcal{K}\hat{\psi} \neq 0$ then, for every fixed $\eta \in (0,1]$, the mapping $\rho \mapsto \|\psi_n^o\|_X$ is continuous and strictly monotonically decreasing with $\lim_{\rho\to\infty}\|\psi_n^o\|_X = 0$. Furthermore, ψ_n^o converges to the solution $\tilde{\psi}_n \in X_n$ of

$$P_n\mathcal{K}^*\big(\eta\alpha^2 + (1-\eta)\beta^2\big)\mathcal{K}\tilde{\psi}_n \;=\; \eta\,P_n\mathcal{K}^*\alpha^2 f^o$$

as ρ tends to zero. From this we conclude that the solution set of (8.47a), (8.47b) again consists of a one-dimensional manifold parametrized by $\eta \in (0,1]$. If $c \geq \|\tilde{\psi}_n\|_X$ the solution (ρ,ψ_n^o) is given by $(\rho,\psi_n^o) = (0,\tilde{\psi}_n)$. If $c < \|\tilde{\psi}_n\|_X$ the solution (ρ,ψ_n^o) is given by the unique solution of (8.47a) and $\|\psi_n^o\|_X = c$.

Example 8.26. In the case of a circular loop we take $X = L^2(0,2\pi)$ and use polar coordinates. Thus it is natural to take Fourier polynomials in order to replace the problem with an approximate one which is finite dimensional. Moreover, we assume that the sets $\mathcal{A}$ and $\mathcal{B}$ are disjoint intervals $[\alpha_1,\alpha_2] \subset [0,2\pi]$ and $[\beta_1,\beta_2] \subset [0,2\pi]$ which correspond to the sections in which we wish to match the given far field pattern and to place nulls, respectively.

We first recall the continuous version of the problem as

$$\begin{array}{ll}\text{Minimize} & \left(\displaystyle\int_{\alpha_1}^{\alpha_2} \big|\mathcal{K}\psi(t) - f^o(t)\big|^2\,dt\,,\ \int_{\beta_1}^{\beta_2} \big|\mathcal{K}\psi(t)\big|^2\,dt\right)^{\top} \\ \text{Subject to} & \psi \in L^2(0,2\pi)\,,\quad \displaystyle\int_0^{2\pi} \big|\psi(s)\big|^2\,ds \leq c^2\,.\end{array} \tag{8.48}$$

Any function $\psi \in L^2(0,2\pi)$ can be represented by its Fourier series

$$\psi(s) \;=\; \sum_{m=-\infty}^{\infty} \psi_m\,\mathrm{e}^{ims}\,,\quad 0 < s < 2\pi\,, \tag{8.49a}$$

with Fourier coefficients

$$\psi_m = \frac{1}{2\pi}\int_0^{2\pi} \psi(s)\,\mathrm{e}^{-ims} ds\,, \quad m \in \mathbb{Z}\,. \tag{8.49b}$$

Therefore, it is appropriate to truncate the series in order to derive the finite dimensional problems. Therefore, defining the $(2n+1)-$dimensional space X_n by $X_n = \operatorname{span}\{\mathrm{e}^{ims} : |m| \leq n\}$ be arrive at the approximated problem

$$\begin{aligned} &\text{Minimize} \quad \left(\int_{\alpha_1}^{\alpha_2} \left|\mathcal{K}\psi(t) - f^o(t)\right|^2 dt\,, \int_{\beta_1}^{\beta_2} \left|\mathcal{K}\psi(t)\right|^2 dt \right)^\top \\ &\text{Subject to } \psi \in X_n\,, \quad \int_0^{2\pi} \left|\psi(s)\right|^2 ds \leq c^2\,. \end{aligned} \tag{8.50}$$

We note that the orthogonal projection operator P_n is just the truncation operator.

For the circular line source of radius a, as in Subsection 7.2.3, the operator $\mathcal{K}$ has the form

$$\mathcal{K}\psi(t) = \int_0^{2\pi} \psi(s)\,\mathrm{e}^{-ika\cos(t-s)}\,ds\,, \quad 0 < t < 2\pi\,. \tag{8.51}$$

As before, we can use the Jacobi-Anger formula

$$\mathrm{e}^{iz\cos\tau} = \sum_{m\in\mathbb{Z}} (-i)^m J_m(z)\,\mathrm{e}^{im\tau}\,,$$

to compute the Fourier series of $\mathcal{K}\psi$ as

$$\mathcal{K}\psi(t) = 2\pi \sum_{m\in\mathbb{Z}} \psi_m\, i^m\, J_m(ka)\,\mathrm{e}^{imt}\,, \quad 0 < t < 2\pi\,. \tag{8.52}$$

Setting $\gamma^2 = \alpha^2 + \eta\beta^2$ as before, we compute

$$\begin{aligned} (\mathcal{K}^*\gamma^2\mathcal{K}\psi)(t) &= \int_0^{2\pi} \gamma^2(s)\,\mathcal{K}\psi(s)\,\mathrm{e}^{ika\cos(t-s)}\,ds \\ &= 2\pi \sum_{m\in\mathbb{Z}} \psi_m\, i^m\, J_m(ka) \int_0^{2\pi} \gamma^2(s)\,\mathrm{e}^{ims}\mathrm{e}^{ika\cos(t-s)}\,ds \end{aligned}$$

and thus, for $\psi \in X_n$,

$$(P_n\mathcal{K}^*\gamma^2\mathcal{K}\psi)(t) = \sum_{|\ell|\leq n} \sum_{|m|\leq n} a_{\ell,m}\,\psi_m\,\mathrm{e}^{imt}$$

where

$$a_{\ell,m} \;=\; 2\pi\, i^{m-\ell}\, J_m(ka)\, J_\ell(ka) \int\limits_0^{2\pi} \gamma^2(s)\, \mathrm{e}^{i(m-\ell)s} ds\,, \quad \ell, m \in \mathbb{Z}\,.$$

Analogously, $\mathcal{K}^*\alpha^2 f^o$ and $P_n\mathcal{K}^*\alpha^2 f^o$ are computed as

$$\begin{aligned}(\mathcal{K}^*\alpha^2 f^o)(t) &= \int\limits_0^{2\pi} \alpha^2(s)\, f^o(s)\, \mathrm{e}^{ika\cos(t-s)}\, ds \\ &= \sum_{m\in\mathbb{Z}} f^o_m\, (-i)^m\, J_m(ka)\, \mathrm{e}^{imt} \int\limits_{\alpha_1}^{\alpha_2} f^o(s)\, \mathrm{e}^{-ims}\, ds\,, \quad 0 \le t \le 2\pi\,,\end{aligned}$$

$$(P_n\mathcal{K}^*\alpha^2 f^o)(t) = \sum_{|m|\le n} f^o_m\, (-i)^m\, J_m(ka)\, \mathrm{e}^{imt} \int\limits_{\alpha_1}^{\alpha_2} f^o(s)\, \mathrm{e}^{-ims}\, ds\,, \quad 0 \le t \le 2\pi\,,$$

respectively. We know from Theorem 8.15 that every weak accumulation point of any sequence $\{\psi^o_n\}$ of Pareto points ψ^o_n of (8.50) is a weak Pareto point of (8.48).

For a numerical simulation we considered again Example 7.6. The desired far field is given by $f^o = \mathcal{K}\hat{\psi}$ which is the optimal solution of the problem to maximize power in the angular section $[0, \pi/4]$. Plots of $|\mathcal{K}\hat{\psi}| = |f^o|$ for wave lengths $\lambda = 1$ and $\lambda = \pi$ are shown in Figure 8.8. We take the interval $[\alpha_1, \alpha_2] = [-\pi/16, \pi/4{+}\pi/16]$ to match f^o and the interval $[\beta_1, \beta_2] = [-\pi/8, 5\cdot \pi/8]$ to place nulls. Figures 8.9 and 8.10 correspond to wave lengths $\lambda = 1$ and $\lambda = \pi$, respectively. We first plot the graphs of $\eta \mapsto |\mathcal{K}\psi^o|$ and then show the patterns $|\mathcal{K}\psi^o|$ for two particular values of η. The first one is small which results in a Pareto point for which the second component is small. The distance to f^o, however, is large. The second value of η is close to one, and we observe that now the first component is small, i.e. $\mathcal{K}\psi^o$ is close to f^o, but there are considerably larger side lobes.

8.4.2 Power and Super-gain

We now turn to the third problem in the table in which we wish to maximize the power in a preassigned sector of the far field while simultaneously minimizing the super-gain. We should recall that we have discussed the notion of the super-gain ratio in Subsections 1.5.2 and 1.5.3 where we considered the linear and circular line sources respectively. We actually proved there that in either

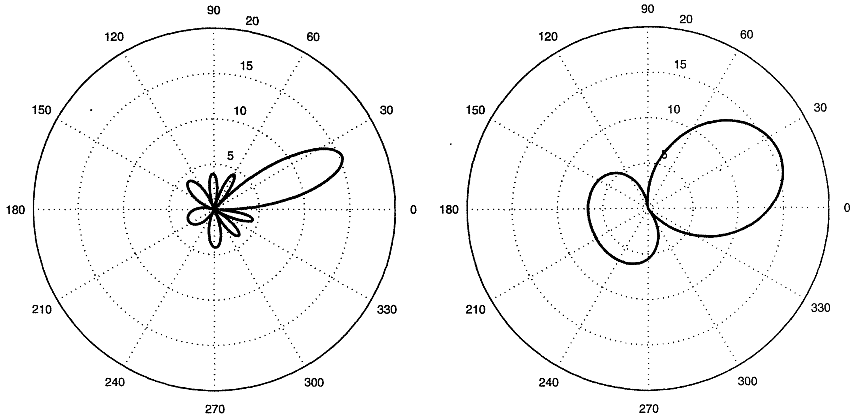


Fig. 8.8. Plots of $|f^o|$ for wave lengths $\lambda = 1$ (left) and $\lambda = \pi$ (right).

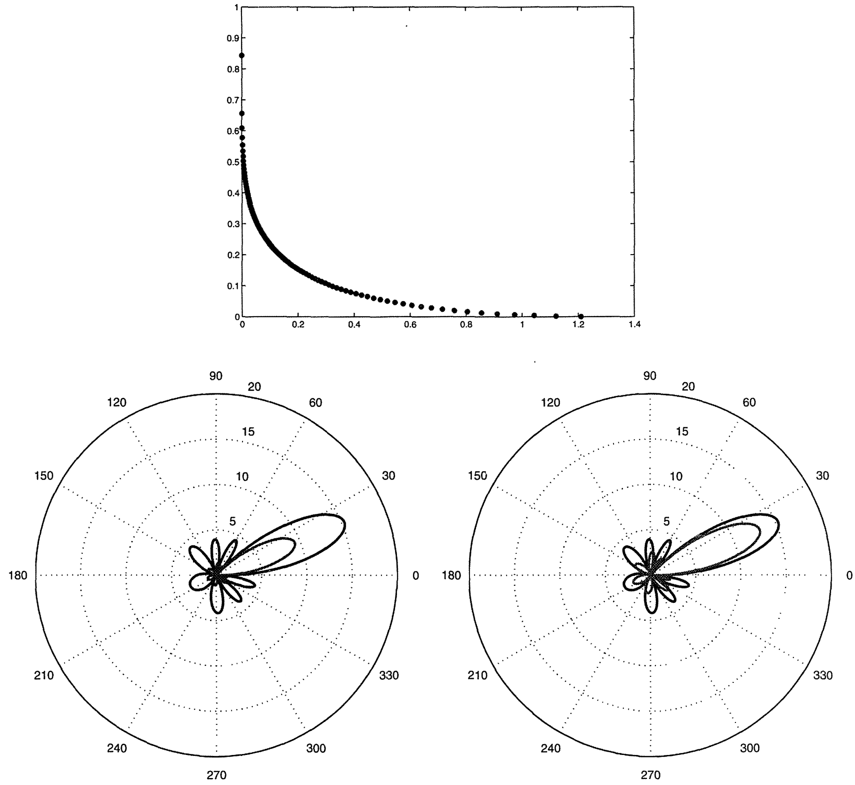


Fig. 8.9. Plots of $\eta \mapsto |\mathcal{K}\psi^o|$ (left) and the far fields $|\mathcal{K}\hat{\psi}|$ and $|\mathcal{K}\psi^o|$ for wave length $\lambda = 1$ corresponding to $\eta = 0.2$ and 0.6, respectively.

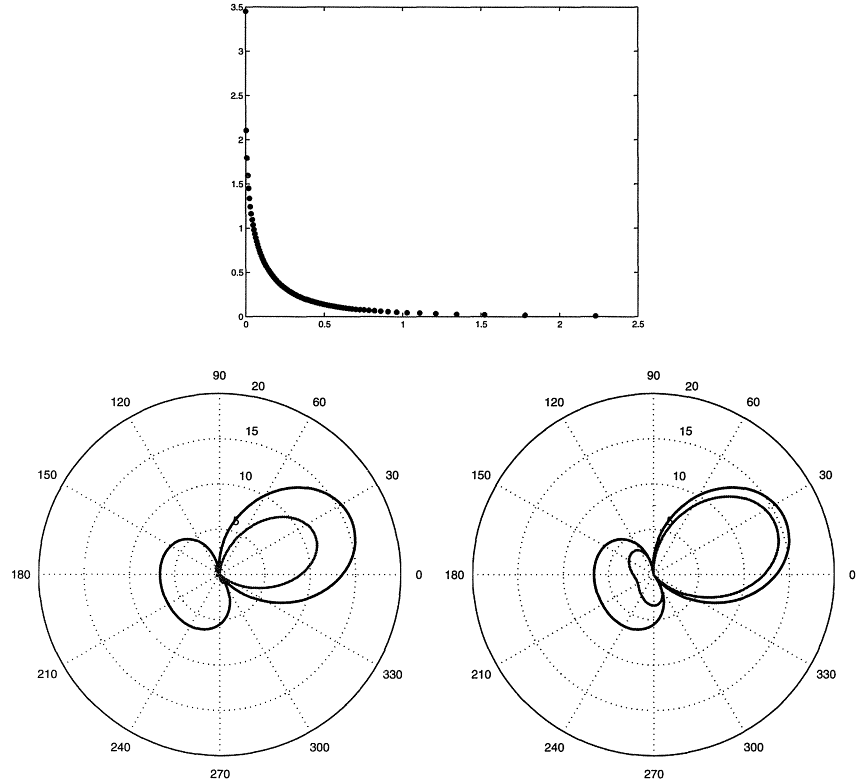


Fig. 8.10. Plots of $\eta \mapsto |\mathcal{K}\psi^o|$ (left) and the far fields $|\mathcal{K}\hat{\psi}|$ and $|\mathcal{K}\psi^o|$ for wave length $\lambda = \pi$ corresponding to $\eta = 0.2$ and 0.9, respectively.

case, the super-gain ratio, $\gamma_\lambda(\psi)$ as defined by Taylor [133] is proportional to the expression given in terms of the far field operator:

$$\frac{\lambda \, \|\psi\|_X^2}{\|\mathcal{K}\psi\|_{L^2(S^{d-1})}^2}, \tag{8.53}$$

where λ is the particular wave-length the proportionality constant being dependent on the particular antenna. In the case of the linear line source, $X = L^2(-\ell, \ell)$, while in the case of the circular line source $X = L^2(S^a)$ where S^a is the circle of radius a.[5] These examples led us, in Chapter 3 (cf. Section 3.4) to introduce the functional

$$\mathcal{J}_6(\psi) = \frac{\|\psi\|_X^2}{\|\mathcal{K}\psi\|_{L^2(S^{d-1})}^2},$$

[5] In particular calculations, we must be careful to remember that $\|\psi\|_{L^2(S^a)} = a\,\|\psi\|_{L^2(0,2\pi)}$

which we called the gain of the antenna. This is something of an abuse of language since, given a particular antenna structure, the super-gain is a function of the distribution ψ.

We should also remember that the idea of a super-gain distribution is a relative one. Any current distribution for which the gain exceeds that of the the gain associated with the uniform distribution $\psi_u \equiv 1$ is called a super-gain antenna. Once again, let us quote Taylor:

> "The answer to the question of how large a value of γ/γ_u can be tolerated in a particular situation will depend on many factors ... . It seems reasonable, however, to look upon source distributions for which γ exceeds γ_u by a factor of ten or more with extreme caution."

The analysis here will show that the use of multi-criteria techniques gives some insight into this comment.

Since we want to control the super-gain ratio while focusing the far field power which are conflicting goals, we can pose a multi-criteria optimization problem as in the entry III of the table.

$$\begin{aligned} &\text{Minimize} \quad \mathcal{F}(\psi) \ := \ \left(-\|\alpha\mathcal{K}\psi\|^2_{L^2(S^{d-1})}\,,\ \frac{\|\psi\|^2_X}{\|\mathcal{K}\psi\|^2_{L^2(S^{d-1})}} \right)^\top \\ &\text{Subject to} \quad \|\psi\|_X \ \leq \ 1, \ \ \psi \neq 0. \end{aligned} \tag{8.54}$$

Our first task is to check that Pareto points exist and to do this, we verify that the hypotheses of part (b) of Theorem 8.8 are satisfied. To this end, let $\mathcal{R} := \{\mathcal{F}(\psi) : 0 < \|\psi\|_X \leq 1\}$, and consider the sets

$$\mathcal{R}_z \ = \ (z - \Lambda) \ \cap \ (\mathcal{R} + \Lambda)\,, \quad z \in \mathbb{R}^2\,, \tag{8.55}$$

where, as usual, Λ is the standard order cone in $\mathbb{R}^2$, i.e.

$$\mathcal{R}_z \ = \ \{\mathcal{F}(\psi) + u : 0 < \|\psi\|_X \leq 1,\ u \in \mathbb{R}^2_{\geq 0}\,,\ \mathcal{F}(\psi) + u \leq z\}.$$

By Theorem 8.8 Pareto points of (8.54) exist provided that for some $z \in \mathbb{R}^2$ the set $\mathcal{R}_z \subset \mathbb{R}^2$ is non-empty and compact. These conditions are proven in the following theorem.

Theorem 8.27. *Assume that the far field operator, $\mathcal{K}$, is compact, one-to-one, and that $\mathcal{K}\psi \in C(S^{d-1})$ is analytic for all $\psi \in X$. Let $\tilde{z} = \mathcal{F}(\tilde{\psi}) \in \mathbb{R}^2$ for some $\tilde{\psi} \in X$ such that $\|\tilde{\psi}\|_X = 1$. Then the set $\mathcal{R}_{\tilde{z}}$ is non-empty and compact and hence the problem (8.54) has Pareto points.*

Proof: First we note that $\tilde{z} = \mathcal{F}(\tilde{\psi}) \in \mathcal{R}_{\tilde{z}}$ i.e., $\mathcal{R}_{\tilde{z}} \neq \emptyset$. To see that this section is bounded, let $z \in \mathcal{R}_{\tilde{z}}$. Then

$$z = \begin{pmatrix} -\|\alpha\mathcal{K}\psi\|^2_{L^2(S^{d-1})} \\ \dfrac{\|\psi\|^2_X}{\|\mathcal{K}\psi\|^2_{L^2(S^{d-1})}} \end{pmatrix} + \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} \le \begin{pmatrix} \tilde{z}_1 \\ \tilde{z}_2 \end{pmatrix}$$

for some $\psi \in X$ such that $0 < \|\psi\|_X \le 1$ and $u_1, u_2 \ge 0$. Hence $z_1 \le \tilde{z}_1$. On the other hand, by the continuity of $\mathcal{K}$, there exists a constant $c > 0$ with $\|\alpha\mathcal{K}\psi\|_{L^2(S^{d-1})} \le c$ for all ψ satisfying $0 < \|\psi\|_X \le 1$, i.e., $-c^2 \le z_1 \le \tilde{z}_1$. Analogously, $0 \le z_2 \le \tilde{z}_2$, and so it is clear that the set $\mathcal{R}_{\tilde{z}}$ is bounded.

In order to show that the set $\mathcal{R}_{\tilde{z}}$ is closed, let z^o be a limit point of a sequence in $\mathcal{R}_{\tilde{z}}$. Then there exist sequences $\{\psi_j\}_{j=1}^\infty \subset X$ with $0 < \|\psi_j\|_X \le 1$ and $\{u_1^{(j)}\}_{j=1}^\infty$, and $\{u_2^{(j)}\}_{j=1}^\infty$ both in $\mathbb{R}_{\ge 0}$ such that

$$\begin{pmatrix} -\|\alpha\mathcal{K}\psi_j\|^2_{L^2(S^{d-1})} + u_1^{(j)} \\ \dfrac{\|\psi_j\|^2_X}{\|\mathcal{K}\psi_j\|^2_{L^2(S^{d-1})}} + u_2^{(j)} \end{pmatrix} \longrightarrow \begin{pmatrix} z_1^o \\ z_2^o \end{pmatrix}.$$

Since $\{\psi_j\}_{j=1}^\infty \subset X$ is bounded, we may assume that this sequence converges weakly to some element, i.e., $\psi_j \rightharpoonup \psi_o$. Moreover, since $\mathcal{K} : X \to L^2(S^{d-1})$ is, by hypothesis, a compact operator we have $\mathcal{K}\psi_j \to \mathcal{K}\psi_o$ in $L^2(S^{d-1})$ as $j \to \infty$. We claim, first, that $\psi_o \ne 0$. To see this, note that $-\|\alpha\mathcal{K}\psi_j\|^2_{L^2(S^{d-1})} + u_1^{(j)} \to z_1^o$. Then $u_1^{(j)} \to u_1^o := z_1^o + \|\alpha\mathcal{K}\psi_o\|^2_{L^2(S^{d-1})} \ge 0$. Hence, since also

$$-\|\alpha\mathcal{K}\psi_j\|^2_{L^2(S^{d-1})} + u_1^{(j)} \le -\|\alpha\mathcal{K}\tilde{\psi}\|^2_{L^2(S^{d-1})}$$

or

$$\|\alpha\mathcal{K}\tilde{\psi}\|^2_{L^2(S^{d-1})} + u_1^{(j)} \le \|\alpha\mathcal{K}\psi_j\|^2_{L^2(S^{d-1})},$$

for all $j = 1, 2, \ldots$, we conclude that

$$\|\alpha\mathcal{K}\psi_o\|^2_{L^2(S^{d-1})} \ge \|\alpha\mathcal{K}\tilde{\psi}\|^2_{L^2(S^{d-1})} + u_1^o \ge \|\alpha\mathcal{K}\tilde{\psi}\|^2_{L^2(S^{d-1})}.$$

By analyticity of $\mathcal{K}\tilde{\psi}$, we may conclude as we have done several times before, $\mathcal{K}\tilde{\psi} \ne 0$ and thus $\psi_o \ne 0$. Note that we have also shown in this argument, that $-\|\alpha\mathcal{K}\psi_j\|^2_{L^2(S^{d-1})} + u_1^{(j)} \to -\|\alpha\mathcal{K}\psi_o\|^2_{L^2(S^{d-1})} + u_1^o$, i.e. $z_1^o = -\|\alpha\mathcal{K}\psi_o\|^2_{L^2(S^{d-1})} + u_1^o$.

Now, consider the second component of $\boldsymbol{\mathcal{F}}$. By choice of z^o, we have that $F_2(\psi_j) + u_2^{(j)} \to z_2^o$ as $j \to \infty$, where the $u_2^{(j)} \ge 0$. We first show that we can find at least a subsequence of the $u_2^{(j)}$ which converges. Indeed, this follows from $0 \le u_2^{(j)} \le F_2(\psi_j) + u_2^{(j)}$ and the convergence $F_2(\psi_j) + u_2^{(j)} \to z_2^o$. Hence, we can find a further subsequence such that (renaming the subsequence if necessary) $u_2^{(j)} \to u_2^o$. Note that this limit $u_2^o \ge 0$.

We now define $\delta \in \mathbb{R}$ by the equation

$$z_2^o \;=\; F_2(\psi_o) \;+\; u_2^o \;+\; \delta\,.$$

It remains to show that $\delta \geq 0$. But this follows from

$$\liminf_{j\to\infty} \|\psi_j\|_X \;\geq\; \|\psi_o\|_X \quad \text{and} \quad \lim_{j\to\infty} \|\mathcal{K}\psi_j\|_{L^2(S^{d-1})} \;=\; \|\mathcal{K}\psi_o\|_{L^2(S^{d-1})}\,,$$

since

$$\delta \;=\; z_2^o - \left[\frac{\|\psi_o\|_X^2}{\|\mathcal{K}\psi_o\|_{L^2(S^{d-1})}^2} + u_2^o\right] \;\geq\; z_2^o - \liminf_{j\to\infty}\left[\frac{\|\psi_j\|_X^2}{\|\mathcal{K}\psi_j\|_{L^2(S^{d-1})}^2} + u_2^{(j)}\right] \;=\; 0\,.$$

It follows that $\boldsymbol{z}^o \in \mathcal{R}_z = (\tilde{\boldsymbol{z}} - \Lambda) \cap (\mathcal{R} + \Lambda)$. This shows that the section $\mathcal{R}_{\tilde{z}}$ is closed. Hence the set $\mathcal{R}_{\tilde{z}}$ is compact in $\mathbb{R}^2$ and therefore there exist Pareto points for the problem (8.54). □

Now we wish to apply the necessary conditions to this problem. Again, the Fréchet derivatives may be calculated using the results in Section 3.4. Those results yield

$$\nabla F_1(\psi) = -2\,\mathcal{K}^*\alpha^2\mathcal{K}(\psi)\,, \tag{8.56a}$$

$$\begin{aligned}\nabla F_2(\psi) &= \frac{2}{\|\mathcal{K}\psi\|_{L^2(S^{d-1})}^4}\left[\|\mathcal{K}\psi\|_{L^2(S^{d-1})}^2\,\psi \;-\; \|\psi\|_X^2\,\mathcal{K}^*\mathcal{K}\psi\right] \\ &= \frac{2}{\|\mathcal{K}\psi\|_{L^2(S^{d-1})}^2}\left[\psi \;-\; Q(\psi)\,\mathcal{K}^*\mathcal{K}\psi\right]\end{aligned} \tag{8.56b}$$

where again $Q(\psi) = \|\psi\|_X^2 / \|\mathcal{K}\psi\|_{L^2(S^{d-1})}^2$.

Now we turn to the application of the Lagrange Multiplier Rule. If we assume that $\psi^o \neq 0$ is a Pareto point for the problem, then the constraint qualification (8.12) is satisfied. The Lagrange Multiplier Rule of Theorem 8.16 guarantees the existence of multipliers, $\eta_1, \eta_2, \rho \geq 0$ with $\eta_1 + \eta_2 > 0$, such that

$$-\eta_1\,\mathcal{K}^*\alpha^2\mathcal{K}\psi^o \;+\; \frac{\eta_2}{\|\mathcal{K}\psi^o\|_{L^2(S^{d-1})}^2}\left[\psi^o \;-\; Q(\psi^o)\mathcal{K}^*\mathcal{K}\psi^o\right] \;+\; \rho\,\psi^o \;=\; 0\,. \tag{8.57}$$

Suppose, first, that $\rho = 0$. Multiplication with ψ^o yields, by the definition of $Q(\psi^o)$,

$$\begin{aligned}0 &= -\eta_1\,\|\alpha\mathcal{K}\psi^o\|_{L^2(S^{d-1})}^2 \;+\; \frac{\eta_2}{\|\mathcal{K}\psi^o\|_{L^2(S^{d-1})}^2}\left[\|\psi^o\|_X^2 - \|\psi^o\|_X^2\right] \\ &= -\eta_1\,\|\alpha\mathcal{K}\psi^o\|_{L^2(S^{d-1})}^2\end{aligned}$$

which implies that $\|\alpha\mathcal{K}\psi^o\|_{L^2(S^{d-1})} = 0$ or $\eta_1 = 0$.

In the former case we see, by the usual arguments, that $\psi^o = 0$ which is a contradiction. Therefore, we must have $\eta_1 = 0$, and this implies that $\eta_2 > 0$ since $\eta_1 + \eta_2 > 0$. Consequently,

$$\left[I - Q(\psi^o)\,\mathcal{K}^*\mathcal{K}\right]\psi^o = 0, \tag{8.58a}$$

where again

$$Q(\psi^o) = \frac{\|\psi^o\|_X^2}{\|\mathcal{K}\psi^o\|_{L^2(S^{d-1})}^2}. \tag{8.58b}$$

We note that (8.58a) is an eigenvalue problem for the self-adjoint operator $\mathcal{K}^*\mathcal{K}$ with eigenvalue $\lambda = 1/Q(\psi^o)$ and eigenfunction ψ^o. If, on the other hand, ψ^o is a normalized eigenfunction of (8.58a) with eigenvalue λ, then we multiply (8.58a) by ψ^o and arrive at

$$\lambda - \|\mathcal{K}\psi^o\|_{L^2(S^{d-1})}^2 = 0,$$

i.e. $1/\lambda = Q(\psi^o)$ and ψ^o solves (8.58a), (8.58b).

Suppose, on the other hand, that $\rho > 0$. Then the inequality constraint is active, i.e. $\|\psi^o\|_X = 1$ and equation (8.57) becomes

$$-\eta_1\,\mathcal{K}^*\alpha^2\mathcal{K}\psi^o + \left(\eta_2\,Q(\psi^o) + \rho\right)\psi^o - \eta_2\,Q(\psi^o)^2\,\mathcal{K}^*\mathcal{K}\psi^o = 0. \tag{8.59}$$

Dividing this equation by $\eta_1 + \eta_2$ and setting $\eta = \eta_2/(\eta_1 + \eta_2)$, we find that the Pareto point must satisfy the system:

$$\left[(1-\eta)\,\mathcal{K}^*\alpha^2\mathcal{K} + \eta\,Q^2\,\mathcal{K}^*\mathcal{K}\right]\psi^o = \left(\frac{\eta_2\,Q + \rho}{\eta_1 + \eta_2}\right)\psi^o, \tag{8.60a}$$

$$\|\psi^o\|_X = 1, \tag{8.60b}$$

and

$$Q = Q(\psi^o) = \frac{1}{\|\mathcal{K}\psi^o\|_{L^2(S^{d-1})}^2}, \tag{8.60c}$$

and these equations serve as the necessary conditions for the Pareto point.

Example 8.28. As a particular example, we consider again the circular line source of Example 7.9 (see also Example 7.6). α is the characteristic function of the interval $[\alpha_1, \alpha_2] \subset [0, 2\pi]$. Then $\mathcal{K}^*\mathcal{K}$ and $\mathcal{K}^*\alpha^2\mathcal{K}$ are given, respectively, by

$$(\mathcal{K}^*\mathcal{K}\psi)(t) = 4\pi^2 \sum_{m\in\mathbb{Z}} J_m(ka)^2\,\psi_m\,\mathrm{e}^{imt}, \quad 0 \le t \le 2\pi,$$

$$(\mathcal{K}^*\alpha^2\mathcal{K}\psi)(t) = \sum_{\ell,m\in\mathbb{Z}} A_{\ell m}\,\psi_m\,\mathrm{e}^{i\ell t}, \quad 0 \le t \le 2\pi,$$

where

$$\psi_m = \frac{1}{2\pi} \int_0^{2\pi} \psi(t)\, e^{-imt}\, dt, \quad m \in \mathbb{Z},$$

are the Fourier coefficients of $\psi \in L^2(0, 2\pi)$ and

$$A_{\ell m} = \begin{cases} 2\pi\, i^{m-\ell}\, J_m(ka)\, J_\ell(ka)\, \dfrac{e^{i(m-\ell)\alpha_2} - e^{i(m-\ell)\alpha_1}}{i\,(m-\ell)}, & \ell \neq m, \\ 2\pi\, J_m(ka)^2\, (\alpha_2 - \alpha_1), & \ell = m. \end{cases}$$

Truncating (8.60a)–(8.60c) results in the following finite dimensional system[6]

$$\left[(1-\eta)\, A + \eta\, Q^2\, D\right] \boldsymbol{x} = \left(\frac{\eta_2\, Q + \rho}{\eta_1 + \eta_2}\right) \boldsymbol{x}, \tag{8.61a}$$

$$\|\boldsymbol{x}\| = 1, \quad \text{and} \quad \boldsymbol{x}^* D \boldsymbol{x} = 1/Q. \tag{8.61b}$$

This is a parametric eigenvalue problem: For every (fixed) parameter $\eta \in [0,1]$ find a parameter $Q > 0$ such that the normalized eigenvector $\boldsymbol{x} = \boldsymbol{x}(Q)$ of (8.61a) satisfies also (8.61b). For the particular case $[\alpha_1, \alpha_2] = [2\pi/9, \pi/3]$ and $\lambda = a$ and the three largest eigenvalues the set of Pareto-critical points are plotted in Figure 8.11. We observe that also values which correspond to the largest eigenvalues are candidates for Pareto optima - a fact which we were not able to prove. The marked points in Figure 8.11 correspond to $\eta = 99/100$ (right mark) and $\eta = 0$ (left mark). Their factors are plotted in Figure 8.12.

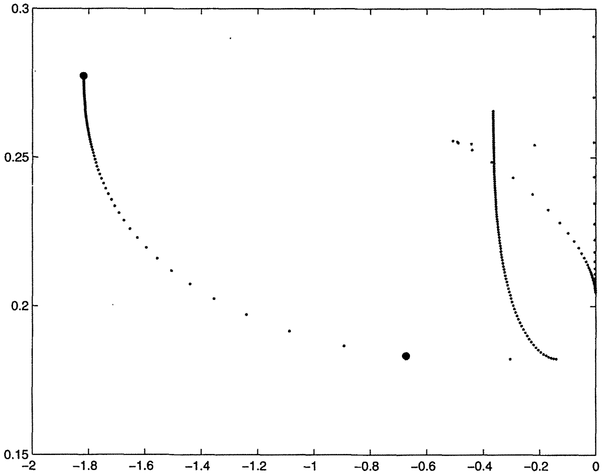


Fig. 8.11. The set of Pareto-critical points of Example 8.28

[6] where we write x instead of ψ^o

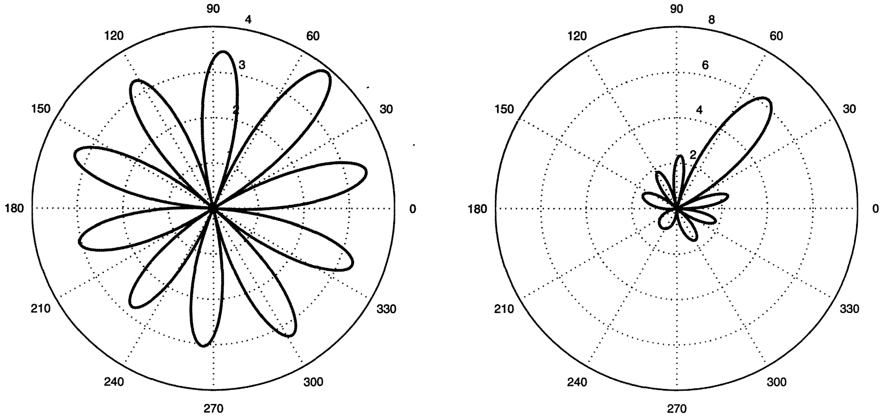


Fig. 8.12. The factors corresponding to the marks in Figure 8.11

8.5 The Signal-to-noise Ratio Problem

8.5.1 Formulation of the Problem and Existence of Pareto Points

We now return to the signal-to-noise ratio problem which we introduced in Section 1.3 and analyzed further in Section 7.4. There, we considered the constrained problem of maximizing the SNR-functional subject to a preassigned bound on the quality factor Q. Now we are prepared to treat the problem as one of multi-criteria optimization.

First, we remind the reader that quality factor, Q, played an important role in the synthesis problem in Chapter 4. One problem treated there was the optimization problem with objective functional

$$\mathcal{J}(\psi) \;=\; \int\limits_{S^{d-1}} \left|(\mathcal{K}\psi)(\hat{x}) - f_0(\hat{x})\right|^2 \, ds \;=\; \|\mathcal{K}\psi - f_0\|^2_{L^2(S^{d-1})}\,, \tag{8.62a}$$

which is to be minimized.

The situation is much as that in the previous section. Good approximations to the desired antenna pattern, f_0, in this mean-square sense can be realized only by producing unacceptable levels of the "quality factor",

$$Q(\psi) \;:=\; \frac{\|\psi\|^2_X}{\|\mathcal{K}\psi\|^2_{L^2(S^{d-1})}}\,. \tag{8.62b}$$

In Section 4.3 we studied the synthesis problem by constraining $Q(\psi)$. In the present context, the appropriate compromises can be studied by identifying the Pareto points for the vector criterion

$$\boldsymbol{\mathcal{F}}(\psi) \;:=\; \begin{pmatrix} \|\mathcal{K}\psi - f_0\|^2_{L^2(S^{d-1})} \\ -\,\|\mathcal{K}\psi\|^2_{L^2(S^{d-1})} \end{pmatrix} \tag{8.63}$$

subject to the power constraint

$$\|\psi\|_X \leq 1. \tag{8.64}$$

In fact, the existence of Pareto points for this problem is guaranteed by the following result:

Theorem 8.29. *The map $\boldsymbol{\mathcal{F}} : X \to \mathbb{R}^2$ is completely continuous and hence Pareto points exist.*

Proof: Since the relatively compact sets in $\mathbb{R}^2$ are the bounded sets, it suffices to check that $\boldsymbol{\mathcal{F}}$ maps bounded sets into bounded sets. But this property of $\boldsymbol{\mathcal{F}}$ follows immediately from the compactness of the operator $\mathcal{K}$. □

Our main interest in this section however is the more difficult problem of optimizing the signal-to-noise ratio (*SNR*) in a given fixed direction $\hat{\boldsymbol{x}} \in S^{d-1}$. We recall that the *SNR* is defined by

$$SNR(\psi) := \frac{|(\mathcal{K}\psi)(\hat{\boldsymbol{x}})|^2}{\int_{S^{d-1}} \omega(\hat{\boldsymbol{y}})^2 \, |\mathcal{K}\psi(\hat{\boldsymbol{y}})|^2 \, ds}, \tag{8.65}$$

where the function $\omega \in L^\infty(S^{d-1})$ is non-zero on a set T of positive measure. The optimization problem studied in §7.4 (see also [72] and [12]) is

$$\text{Maximize} \quad SNR(\psi) \quad \text{subject to} \quad \|\psi\|_X \leq 1 \text{ and } Q(\psi) \leq c, \tag{8.66}$$

where $Q(\psi)$ is given by (8.62b), and $c > 0$ is a fixed constant.

We make the same assumptions as at the beginning of Chapter 7, namely that

(A1) $\mathcal{K} : X \to C(S^{d-1})$ is compact and one-to-one. In particular, $\mathcal{K}$ is not identically zero.

(A2) $\mathcal{K}\psi \in C(S^{d-1})$ is an analytic function on S^{d-1} for every $\psi \in X$.

However, we now consider, not the constrained problem (8.66), but rather the vector valued optimization problem

$$\text{Minimize} \quad \boldsymbol{\mathcal{F}}(\psi) := \begin{pmatrix} -SNR(\psi) \\ Q(\psi) \end{pmatrix} \quad \text{subject to} \quad \psi \neq 0. \tag{8.67}$$

In order to prove that Pareto points for the problem (8.67) exist we will again use Theorem 8.10. This requires that, for some $\boldsymbol{z} \in \mathbb{R}^2$, the set

$$S_z := \left\{ \boldsymbol{\mathcal{F}}(\psi) + \boldsymbol{u} \in \mathbb{R}^2 : \psi \neq 0, \ \boldsymbol{u} \geq 0, \ \boldsymbol{\mathcal{F}}(\psi) + \boldsymbol{u} \leq \boldsymbol{z} \right\} \tag{8.68}$$

is compact in $\mathbb{R}^2$. First, we show that S_z is bounded. Assume on the contrary that there exist sequences $\{\psi_j\} \subset X$ and $\{\boldsymbol{u}^{(j)}\} \subset \mathbb{R}^2_{\geq 0}$ with $\psi_j \neq 0$ and $\boldsymbol{\mathcal{F}}(\psi_j) + \boldsymbol{u}^{(j)} \leq \boldsymbol{z}$ and $SNR(\psi_j) + u_1^{(j)} \to -\infty$ as j tends to infinity. (Note that always $Q(\psi_j) + u_2^{(j)} \geq 0$.) Since $\mathcal{F}$ is scale invariant i.e. for any scalar $\rho \in \mathbb{C} \setminus \{0\}$ we have $\mathcal{F}(\rho\psi) = \mathcal{F}(\psi)$, we can assume that $\|\psi_j\|_X = 1$ and thus $\{\psi_j\}$ contains a weak limit point. Without loss of generality we assume that $\psi_j \rightharpoonup \psi$ weakly in X for some $\psi \in X$. From $Q(\psi_j) = 1/\|\mathcal{K}\psi_j\|_X^2 \leq z_2$ and the compactness of $\mathcal{K}$ we conclude that $\mathcal{K}\psi \neq 0$ and thus $\psi \neq 0$. Since $\mathcal{K}$ is also compact as a map into $C(S^{d-1})$ we have $(\mathcal{K}\psi_j)(\hat{\boldsymbol{x}}) \to (\mathcal{K}\psi)(\hat{\boldsymbol{x}})$ and thus $SNR(\psi_j) \to SNR(\psi)$. This contradicts the assumption that $SNR(\psi_j) + u_1^{(j)} \to -\infty$.

By essentially the same arguments we can show that S_z is closed. Indeed, let again $\{\psi_j\} \subset X$, $\{\boldsymbol{u}^{(j)}\} \subset \mathbb{R}^2_{\geq 0}$ with $\psi_j \neq 0$ and $\mathcal{F}(\psi_j) + \boldsymbol{u}^{(j)} \leq \boldsymbol{z}$ and $\mathcal{F}(\psi_j) + \boldsymbol{u}^{(j)} \to \boldsymbol{z}^o \in \mathbb{R}^2$. Again, we can assume that $\|\psi_j\|_X = 1$ and thus $\psi_j \rightharpoonup \psi$ weakly in X for some $\psi \in X$ with $\|\psi\|_X \leq 1$. The facts that $\psi \neq 0$ and $SNR(\psi_j) \to SNR(\psi)$ follow as above. Therefore, also $\{u_1^{(j)}\}$ contains an accumulation point, i.e, without loss of generality $u_1^{(j)} \to u_1$.

Finally, we set $u_2 := z_2 - Q(\psi)$ and have to show that $u_2 \geq 0$. This follows from

$$\begin{aligned} u_2 = z_2 \; - \; \frac{\|\psi\|_X^2}{\|\mathcal{K}\psi\|^2_{L^2(S^{d-1})}} \; &\geq \; z_2 \; - \; \frac{1}{\|\mathcal{K}\psi\|^2_{L^2(S^{d-1})}} \\ = z_2 \; - \; \lim_{j\to\infty} \frac{1}{\|\mathcal{K}\psi_j\|^2_{L^2(S^{d-1})}} \; & z_2 \; - \; \lim_{j\to\infty} Q(\psi_j) \; \geq \; 0 \end{aligned}$$

which means that $\boldsymbol{z}^o \in S_z$. Hence the set S_z is closed. Application of Theorem 8.10 yields:

Corollary 8.30. *Under the assumption of the previous theorem there exist Pareto points of (8.67).*

8.5.2 The Lagrange Multiplier Rule

Now we will apply Theorem 8.16 to the optimization problem (8.67) and will use the resulting equations to compute the set of all "critical points" which, as in the case of a single cost functional, contains the set of Pareto points. Much of the necessary work has already been done. Indeed, we already have computed the Fréchet derivatives of *SNR* and Q at $\psi^o \in X$ in Section 3.4. Using the Riesz representation $p \in X$ of the functional $\psi \mapsto (\mathcal{K}\psi)(\hat{\boldsymbol{x}})$, i.e.

$$(\mathcal{K}\psi)(\hat{\boldsymbol{x}}) \;=\; (\psi, p)_X \quad \text{for all } \psi \in X\,,$$

we can write the gradients in the form

$$\begin{aligned}\nabla SNR(\psi^o) = {} & \frac{2}{\|\omega\mathcal{K}\psi^o\|^4_{L^2(S^{d-1})}}\Big[\|\omega\mathcal{K}\psi^o\|^2_{L^2(S^{d-1})}\,(\psi^o,p)_X\,p \\ & - \,|(\mathcal{K}\psi^o)(\hat{\boldsymbol{x}})|^2\,\mathcal{K}^*\omega^2\mathcal{K}\psi^o\Big], \\ \nabla Q(\psi^o) = {} & \frac{2}{\|\mathcal{K}\psi^o\|^4_{L^2(S^{d-1})}}\Big[\|\mathcal{K}\psi^o\|^2_{L^2(S^{d-1})}\,\psi^o \;-\; \|\psi^o\|^2_X\,\mathcal{K}^*\mathcal{K}\psi^o\Big].\end{aligned}$$

Let ψ^o be a Pareto point. Application of Theorem 8.16 yields the existence of multipliers $\eta_1,\eta_2 \geq 0$ with $\eta_1+\eta_2>0$ and $-\eta_1\,\nabla SNR(\psi^o) + \eta_2\,\nabla Q(\psi^o) = 0$, i.e.

$$\begin{aligned}&-\frac{\eta_1}{\|\omega\mathcal{K}\psi^o\|^4_{L^2(S^{d-1})}}\Big[\|\omega\mathcal{K}\psi^o\|^2_{L^2(S^{d-1})}\,(\psi^o,p)_X\,p - |(\mathcal{K}\psi^o)(\hat{\boldsymbol{x}})|^2\,\mathcal{K}^*(\omega^2\mathcal{K}\psi^o)\Big] + \\ &+\frac{\eta_2}{\|\mathcal{K}\psi^o\|^4_{L^2(S^{d-1})}}\Big[\|\mathcal{K}\psi^o\|^2_{L^2(S^{d-1})}\,\psi^o - \|\psi^o\|^2_{L^2(S^{d-1})}\,\mathcal{K}^*\mathcal{K}\psi^o\Big] \;=\; 0,\end{aligned}$$

or

$$\begin{aligned}&\eta_1\,\frac{|(\mathcal{K}\psi^o)(\hat{\boldsymbol{x}})|^2}{\|\omega\mathcal{K}\psi^o\|^4_{L^2(S^{d-1})}}\,\mathcal{K}^*(\omega^2\mathcal{K}\psi^o) \;+\; \frac{\eta_2}{\|\mathcal{K}\psi^o\|^2_{L^2(S^{d-1})}}\psi^o \;- \\ &-\;\frac{\eta_2}{\|\mathcal{K}\psi^o\|^2_{L^2(S^{d-1})}}Q_0\mathcal{K}^*\mathcal{K}\psi^o \;=\; \frac{\eta_1(\psi^o,p)_X}{\|\omega\mathcal{K}\psi^o\|^2_{L^2(S^{d-1})}}\,p\end{aligned}$$

where $Q_0 = \|\psi^o\|^2_X \,/\, \|\mathcal{K}\psi^o\|^2_{L^2(S^{d-1})}$.

Now we distinguish between two cases:

Case 1: $\eta_1\,(\mathcal{K}\psi^o)(\hat{\boldsymbol{x}}) = 0$. Then $\mathcal{K}^*\mathcal{K}\psi^o = \frac{1}{Q_0}\,\psi^o$, i.e. $1/Q_0$ is an eigenvalue of $\mathcal{K}^*\mathcal{K}$ with eigenfunction ψ^o.

Case 2: $\eta_1 > 0$ and $(\mathcal{K}\psi^o)(\hat{\boldsymbol{x}}) \neq 0$. We then set

$$\eta \;:=\; \frac{\eta_2}{\eta_1\,\|\mathcal{K}\psi^o\|^2_{L^2(S^{d-1})}}\;\frac{\|\omega\mathcal{K}\psi^o\|^4_{L^2(S^{d-1})}}{|(\mathcal{K}\psi^o)(\hat{\boldsymbol{x}})|^2} \quad\text{and}\quad \psi^{oo} \;:=\; \frac{\overline{(\mathcal{K}\psi^o)(\hat{\boldsymbol{x}})}}{\|\omega\mathcal{K}\psi^o\|^2_{L^2(S^{d-1})}}\,\psi^o.$$

Then the function ψ^{oo} is also Pareto optimal and $(\mathcal{K}\psi^{oo})(\hat{\boldsymbol{x}}) = \|\omega\mathcal{K}\psi^{oo}\|^2_{L^2(S^{d-1})}$, thus $SNR(I_1) = \|\omega\mathcal{K}\psi^{oo}\|^2_{L^2(S^{d-1})}$, and

$$\mathcal{K}^\star(\omega^2\mathcal{K}\psi^{oo}) \;+\; \eta\,\psi^{oo} \;-\; \eta\,Q_0\,\mathcal{K}^*\mathcal{K}\psi^{oo} \;=\; p \quad \text{in}\,X\,. \tag{8.69}$$

Therefore we see that if ψ^{oo} is a Pareto point of (8.67) which is normalized so that $(\mathcal{K}\psi^{oo})(\hat{\boldsymbol{x}}) = \|\omega\mathcal{K}\psi^{oo}\|^2_{L^2(S^{d-1})}$ then there exists $\eta \geq 0$ with (8.69) where $Q_0 = \|\psi^{oo}\|^2_X \,/\, \|\mathcal{K}\psi^{oo}\|^2_{L^2(S^{d-1})}$.

If, on the other hand, ψ^{oo} solves (8.69) for some Q_0 and $\eta > 0$ then

$$\begin{aligned}(\mathcal{K}\psi^{oo})(\hat{\boldsymbol{x}}) &= (\psi^{oo},p)_X \\ &= \|\omega\mathcal{K}\psi^{oo}\|^2_{L^2(S^{d-1})} \;+\; \eta\,\|\psi^{oo}\|^2_X \;-\; \eta\,Q_0\;\|\mathcal{K}\psi^{oo}\|^2_{L^2(S^{d-1})}\,,\end{aligned}$$

i.e,

$$(\mathcal{K}\psi^{oo})(\hat{\boldsymbol{x}}) \;-\; \|\omega\mathcal{K}\psi^{oo}\|^2_{L^2(S^{d-1})} \;=\; \eta\left[\|\psi^{oo}\|^2_X \;-\; Q_0\,\|\mathcal{K}\psi^{oo}\|^2_{L^2(S^{d-1})}\right].$$

Hence

$$(\mathcal{K}\psi^{oo})(\hat{\boldsymbol{x}}) \;=\; \|\omega\mathcal{K}\psi^{oo}\|^2_{L^2(S^{d-1})} \tag{8.70a}$$

is equivalent to

$$\|\psi^{oo}\|^2_X \;=\; Q_0\,\|\mathcal{K}\psi^{oo}\|^2_{L^2(S^{d-1})}\,. \tag{8.70b}$$

Equations (8.69) and (8.70) describe a one-parameter family of critical points which contains the set of Pareto points as well as the weak Pareto points.

8.5.3 An Example

We illustrate this approach which uses the necessary optimality conditions with a numerical example (c.f. Section 7.4 for the related example and numerical results for the problem where Q is fixed and SNR is to be maximized). We consider the case where ω is the characteristic function of a portion of the unit circle, e.g. if $0 \le t_1 < t_2 \le 2\pi$,

$$\omega(t) \;=\; \begin{cases} 1, \text{ if } t_1 \le t \le t_2\,, \\ 0, \text{ otherwise.} \end{cases}$$

As an example, we take for $\mathcal{K}$ the particular far field operator for the circular line source of radius 1. Let (θ,ϕ) are the spherical polar coordinates of $\hat{\boldsymbol{x}}$ and

$$(K\psi)(\theta) \;:=\; \int\limits_0^{2\pi} \psi(s)\,\mathrm{e}^{-iks\cos\theta}ds\,, \quad 0 \le \theta \le 2\pi\,. \tag{8.71}$$

Then $p(s) = \exp(iks\cos\theta_0)$ is the Riesz representation of $\psi \mapsto (K\psi)(\hat{\boldsymbol{x}}_0)$. Let

$$\psi(s) \;:=\; \sum_{j=-\infty}^{\infty} x_j\,\mathrm{e}^{ijs}\,.$$

Then

$$(K\psi)(t) \;=\; \sum_{j=-\infty}^{\infty} x_j \int\limits_0^{2\pi} \mathrm{e}^{ijs}\,\mathrm{e}^{-ik\cos(t-s)}\,ds \;=\; 2\pi \sum_{j=-\infty}^{\infty} x_j\,(-i)^j\,J_j(k)\,\mathrm{e}^{ijt}\,,$$

where we have used the Jacobi-Anger expansion (see ([90]))

$$\mathrm{e}^{-ik\cos\tau} \;=\; \sum_{n=-\infty}^{\infty} (-i)^n\,J_n(k)\,\mathrm{e}^{in\tau}\,.$$

The operator K can be represented as an infinite diagonal matrix with elements $2\pi\,(-i)^j\,J_j(k)$, K^*K is diagonal with elements $d_j := 4\pi^2 J_j(k)^2$ and

$$\begin{aligned} K^*(\omega^2 K\psi)(t) &= 2\pi \sum_{j=-\infty}^{\infty} x_j\,(-i)^j\,J_j(k) \int_{t_1}^{t_2} \mathrm{e}^{ijs}\,\mathrm{e}^{ik\cos(t-s)}\,ds \\ &= \sum_{\ell=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} a_{\ell j}\,x_j\,\mathrm{e}^{i\ell t} \end{aligned}$$

where the coefficients $a_{\ell j}$ are defined by

$$\begin{aligned} a_{\ell j} &= 2\pi\,(-i)^j\,J_j(k)\,i^\ell\,J_\ell(k) \int_{t_1}^{t_2} \mathrm{e}^{i(j-\ell)s}\,ds \\ &= \begin{cases} 2\pi\,i^{\ell-j}\,J_j(k)\,J_\ell(k)\,\frac{1}{i(j-\ell)}\,\left(\mathrm{e}^{i(j-\ell)t_2} - \mathrm{e}^{i(j-\ell)t_1}\right), & \text{if } j \neq \ell, \\ 2\pi\,J_j(k)^2\,(t_2 - t_1), & \text{if } j = \ell. \end{cases} \end{aligned}$$

We project equation (8.69) onto the finite dimensional space $X_n = \operatorname{span}\{\mathrm{e}^{ijt} : |j| \le n\}$. Then equations (8.69) and (8.70) take the discretized form

$$(\boldsymbol{A} + \tilde{\eta}\boldsymbol{I} - \tilde{\eta}\,Q_0\,\boldsymbol{D})\,\boldsymbol{x} \;=\; \boldsymbol{r} \quad \text{and} \quad \|\boldsymbol{x}\|^2 - Q_0\boldsymbol{x}^*\boldsymbol{D}\boldsymbol{x} \;=\; 0 \tag{8.72}$$

where $\boldsymbol{I}$ is the identity matrix,

$$\begin{aligned} \boldsymbol{D} &:= \operatorname{diag}(d_j : |j| \le n) \quad \text{with} \quad d_j \;:=\; 4\pi^2\,J_j(k)^2\,, \\ \boldsymbol{A} &:= (a_{\ell j})_{\ell,j=-n,\dots,n}, \quad \text{and} \quad r_j \;:=\; i^j\,J_j(k)\,\mathrm{e}^{-ij\theta_0}\,. \end{aligned}$$

We have carried out specific computations for this example with the specific choice of parameters $k = 6$, $t_1 = 40^o$, $t_2 = 140^o$, $\theta_0 = 90^o$ and $n = 32$. For large ranges of η (from .1 to 10) we computed, by a simple bisection method, all zeros of the function $\varphi(Q_0) := \|\boldsymbol{x}\|^2/\big(\boldsymbol{x}^*\boldsymbol{D}\boldsymbol{x}\big) - Q_0$, where $\boldsymbol{x}$ solves the first equation of (8.72) for Q_0. The numerical results show that, for every fixed value of η, the function $\varphi(Q_0)$ has several zeros which correspond to local Pareto points. In Figure 8.13, left, for 100 values of η between 0.1 and 10 we marked the values $\big(-\mathit{SNR}(\boldsymbol{x}), Q(\boldsymbol{x})\big)$ for the smallest and the second smallest zero of φ. The right plot of Figure 8.13 shows a section of the same plot.

The following table lists some values of η ranging between 0.1 and 10, together with corresponding values of *SNR* and Q_0.

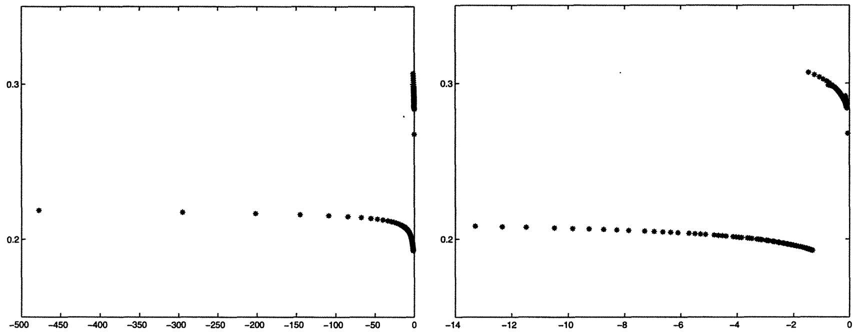


Fig. 8.13. Pareto points (lower branch) and other critical points (upper branch)

$\tilde{\eta}$	*SNR*	Q_0	$\tilde{\eta}$	*SNR*	Q_0
0.1	478	0.219	5.0	3.24	0.200
0.5	109	0.215	5.5	2.89	0.199
1.0	40.2	0.213	6.0	2.53	0.198
1.5	20.8	0.210	6.5	2.24	0.197
2.0	13.3	0.208	7.0	2.09	0.196
2.5	9.27	0.207	7.5	1.90	0.196
3.0	6.95	0.205	8.0	1.74	0.195
3.5	5.49	0.203	8.5	1.61	0.194
4.0	4.52	0.202	9.0	1.50	0.194
4.5	3.84	0.201	10.	1.33	0.193

As we have seen above, the value of Q_0 is bounded below by $1/\mu_{max}$ where μ_{max} is the largest eigenvalue of $\mathcal{K}^*\mathcal{K}$. In our example $1/\mu_{max} = 0.1822$.

We note, finally, that this lower branch shows relatively wide variation in the value of SNR for very small changes in the value of the quality factor Q. This indicates that one should be able to achieve relatively high values of SNR without an appreciable degradation of the quality factor.

A

Appendix

A.1 Introduction

In this appendix we have collected some mathematical facts that we believe will be useful to the reader; a quick summary of some facts from real and functional analysis that we have used in this monograph. We would, of course, like to make this book as self-contained as possible. It is not possible to give a complete overview, with or without proofs, of all the mathematical facts that we have used; to do so would require another volume. There are plenty of very good sources to which we can refer the reader in full confidence that those books contain all the necessary details and are clearly written.

In particular, while not intending to ignore many excellent texts, we have two classics in mind. First, the engineering community has been fortunate to have available, since the early 1970's, the book of A. Naylor and G. Sell [108]. It has the advantage of providing many concrete engineering examples to illustrate the application of the different topics. The other, is the now classic book of David G. Luenberger [88]. Luenberger's book has been used by generations of students, in disparate fields, who have need of a good foundation in functional analysis. For those who simply want to look up a particular result, we have also cited the books of Yosida [145] and the first volume of Dunford and Schwartz [36].

So we will try, mostly by examples, to remind the reader of some of the basic facts that form the basis of our exposition. There are many spots in the text that refer specifically to the appendix for a precise formulation of some results; those will all be found here. As we progress to less standard material, we will put in more details. Indeed, the final portion of the appendix is devoted to ordered vector spaces and Pareto optimality, a subject that does not commonly appear in the basic texts. There we provide more detail.

It is our hope that the material gathered here will substantially aid whose who have been so kind as to look into our book.

A.2 Basic Notions and Examples

We start by assuming that the reader is familiar with the basic definition of vector spaces over the real and complex fields, and that the notions of norm and inner product on a vector space are likewise familiar. We call the elements in the vector space **vectors**.

In a **pre-Hilbert space**, that is, in a vector space with an inner product which is, by definition, homogeneous with respect to the first argument, we recall that the inner product is associated in a natural way with a norm according to the equation $(x, x) = \|x\|^2$.

Of course, $\mathbb{R}^n$ and $\mathbb{C}^n$ are examples of pre-Hilbert spaces, but our main concern is with infinite dimensional function spaces. Let us give some quick examples. In all of them we use $\mathbb{F}$ to stand for either $\mathbb{R}$ or $\mathbb{C}$.

Examples A.1.

(a) Let $C[a, b]$ be the space of all continuous $\mathbb{F}$-valued functions defined on the closed bounded interval $[a, b] \subset \mathbb{R}$. Then $C[a, b]$ is a vector space over $\mathbb{F}$. Moreover, we can define a norm on this space by

$$\|x\|_\infty := \sup_{a \le t \le b} |x(t)|, \quad \text{for } x \in C[a, b]. \tag{A.1}$$

(b) Look at the same vector space, but this time, introduce an inner product on the space by taking

$$(x, y)_{L^2} = \int_a^b x(t)\,\overline{y(t)}\,dt, \quad x, y \in C[a, b]. \tag{A.2a}$$

Then $C[a, b]$ is a pre-Hilbert space and the inner product then defines the corresponding norm which we write as

$$\|x\|_{L^2} = \sqrt{(x, x)_{L^2}} = \sqrt{\int_a^b |x(t)|^2\,dt}, \quad x \in C[a, b]. \tag{A.2b}$$

(c) Let $m \in \mathbb{N}$ and $\alpha \in (0, 1]$. We define the spaces $C^m[a, b]$ and $C^{m,\alpha}[a, b]$ by

$$C^m[a, b] := \{x \in C[a, b] : x^{(k)} \in C[a, b],\ 1 \le k \le m\} \tag{A.3a}$$

and

$$C^{m,\alpha}[a, b] := \left\{ x \in C^m[a, b] : \sup_{t \ne s} \frac{|x^{(m)}(t) - x^{(m)}(s)|}{|t - s|^\alpha} < \infty \right\} \tag{A.3b}$$

and we equip them, respectively, with norms

$$\|x\|_{C^m} \;:=\; \max_{0\le k\le m} \left\|x^{(k)}\right\|_\infty , \tag{A.3c}$$

and

$$\|x\|_{C^{m,\alpha}} \;:=\; \|x\|_{C^m} \;+\; \sup_{s\neq t} \frac{\left|x^{(m)}(t) - x^{(m)}(s)\right|}{|t-s|^\alpha} . \tag{A.3d}$$

Here we denote by $x^{(k)}$ the k^{th} derivative of the function x.

(d) We give here an example of a vector space of infinite sequences which we will call c_0 over $\mathbb{F}$. This space is defined by

$$c_0 \;:=\; \left\{ \{x_k\}_{n=1}^\infty \subset \mathbb{F} : \lim_{k\to\infty} x_k = 0 \right\} , \tag{A.4a}$$

and is a normed space with respect to the norm

$$\|x\| \;:=\; \sup_{k\in\mathbb{N}} |x_k| . \tag{A.4b}$$

(e) Finally, we mention the vector spaces of p-summable sequences in $\mathbb{F}$ which are denoted by ℓ^p. We will restrict ourselves to indices $1 \le p \le \infty$. These spaces are defined by

$$\ell^p \;:=\; \left\{ \{x_k\}_{n=1}^\infty \subset \mathbb{F} : \sum_{n=1}^\infty |x_k|^p < \infty \right\} , \quad \text{for } 1 \le p < \infty , \tag{A.5a}$$

for which we define the norm as

$$\|x\|_p \;:=\; \left(\sum_{k=1}^\infty |x_k|^p \right)^{1/p} , \tag{A.5b}$$

and in the case $p = \infty$, by

$$\ell^\infty \;:=\; \left\{ \{x_k\}_{n=1}^\infty \subset \mathbb{F} : \{x_k\}_{k=1}^\infty \text{ is bounded } \right\} , \tag{A.5c}$$

equipped with the norm

$$\|x\|_\infty \;:=\; \sup_{k\in\mathbb{N}} |x_k| . \tag{A.5d}$$

Note that the space $c_0 \subset \ell^\infty$. We note that these spaces are all normed spaces but that ℓ^2 is a pre-Hilbert space when we define an inner product by

$$(x,y)_{\ell^2} \;:=\; \sum_{k=1}^\infty x_k\, \overline{y}_k , \quad \text{where } x = \{x_k\}_{k=1}^\infty ,\; y = \{y_k\}_{k=1}^\infty . \tag{A.5e}$$

Once we have a norm, then we can define the open and closed balls with radius r and center $x_o \in X$, respectively, by

$$B(x_o, r) := \{x \in X : \|x - x_o\| < r\}, \quad \text{and} \tag{A.6a}$$

$$B[x_o, r] := \{x \in X : \|x - x_o\| \leq r\}. \tag{A.6b}$$

These definitions make it easier to define some standard topological ideas as, for example, that of bounded set or those of closed or open sets and of convergence.

Definition A.2. *Let X be a normed space over the field $\mathbb{F} = \mathbb{R}$ or $\mathbb{C}$.*

(a) A subset $M \subset X$ is called **bounded** *if there exists $r > 0$ with $M \subset B(x, r)$. The set $M \subset X$ is called* **open** *if for every $x \in M$ there exists $\epsilon > 0$ such that $B(x, \epsilon) \subset M$. The set $M \subset X$ is called* **closed** *if the complement $X \setminus M$ is open.*

(b) A sequence $\{x_k\}_{k=1}^{\infty} \subset X$ is called **bounded** *if there exists $c > 0$ such that $\|x_k\| \leq c$ for all k. The sequence $\{x_k\}_{k=1}^{\infty} \subset X$ is called* **norm convergent** *if there exists $x \in X$ such that $\|x - x_k\|$ converges to zero in $\mathbb{R}$. We denote the limit by $x = \lim_{k \to \infty} x_k$, or we write $x_k \to x$ as $k \to \infty$. The sequence $\{x_k\}_{k=1}^{\infty} \subset X$ is called a* **Cauchy sequence** *if for every $\epsilon > 0$ there exists $N \in \mathbb{N}$ with $\|x_m - x_k\| < \epsilon$ for all $m, k \geq N$.*

(c) Let $\{x_k\}_{k=1}^{\infty} \subset X$ be a sequence. $x \in X$ is called an **accumulation point** *if there exists a subsequence $\{a_{k_n}\}_{n=1}^{\infty}$ that converges to x.*

(d) A set $M \subset X$ is called **compact** *if every sequence in M has an accumulation point in M.*

We can make these definitions concrete by the following specific example.

Example A.3. Let $X = C[0, 1]$ over $\mathbb{R}$ and $x_k(t) = t^k$, $t \in [0, 1]$, $k \in \mathbb{N}$. The sequence $\{x_k\}_{k=1}^{\infty}$ converges to zero with respect to the Euclidean norm $\|\cdot\|_{L^2}$ introduced in (A.2b). With respect to the supremum norm $\|\cdot\|_{\infty}$ of (A.1), however, the sequence does not converge to zero since $\|x_k\|_{\infty} = 1$ for all k. In fact, the sequence does not even have an accumulation point. This shows that the closed, bounded set $B[0, 1] = \{x \in C[0, 1] : \|x\|_{\infty} \leq 1\}$ is *not* compact.

It is easy to prove that a set M is closed if and only if the limit of every convergent sequence $\{x_k\}_{k=1}^{\infty} \subset M$ also belongs to M.

Definition A.4.

(a) The sets

$$\operatorname{int} M = \overset{o}{M} := \{x \in M : \text{there exists } \epsilon > 0 \text{ with } B(x, \epsilon) \subset M\}$$

and

$$c\ell\, M = \overline{M} := \{x \in X : \exists \{x_k\}_{k=1}^{\infty} \subset M \text{ with } x = \lim_{k \to \infty} x_k\}$$

are called the **interior** *and* **closure**, *respectively, of M.*

(b) The set $M \subset X$ is called **dense** *in X if $\overline{M} = X$.*

As we have seen in Example A.3, in general the topological properties of a set depend on the norm in X, except in the case of finite dimensional spaces where there is essentially only one norm topology or, put another way, only one notion of convergence. More precisely, one can show that, in either $\mathbb{R}^n$ or $\mathbb{C}^n$, if $\|\cdot\|_1$ and $\|\cdot\|_2$ are two given norms, then these norms are equivalent in the sense that there exist constants $c_2 \geq c_1 > 0$ with

$$c_1 \|x\|_1 \;\leq\; \|x\|_2 \;\leq\; c_2 \|x\|_1 \quad \text{for all } x \in \mathbb{F}^n .$$

In other words, every ball with respect to $\|\cdot\|_1$ contains a ball with respect to $\|\cdot\|_2$ and vice versa. Further properties are collected in the following theorem.

Theorem A.5. *Let X be a normed space over $\mathbb{F}$ and $M \subset X$ be a subset.*

(a) M is closed if and only if $M = \overline{M}$, and M is open if and only if $M = \overset{o}{M}$.
(b) If $M \neq X$ is a linear subspace, then $\overset{o}{M} = \emptyset$, and $\overline{M}$ is also a linear subspace.
(c) In finite dimensional spaces, every subspace is closed.
(d) Every compact set is closed and bounded. In finite dimensional spaces, the reverse is also true (Theorem of Bolzano-Weierstrass): In a finite dimensional normed space, every closed and bounded set is compact.

A crucial property of the set of real numbers is its *completeness.* It is also a necessary assumption for many results in functional analysis.

Definition A.6. *(Banach Space, Hilbert Space)*

A normed space X over $\mathbb{F}$ is called **complete** *or a* **Banach space** *if every Cauchy sequence converges in X. A complete pre-Hilbert space is called a* **Hilbert space**.

The spaces $\mathbb{C}^n$ and $\mathbb{R}^n$ are Hilbert spaces with respect to their canonical inner products. The space $C[a,b]$ is *not* complete with respect to the inner product $(\cdot,\cdot)_{L^2}$ of (A.2a)! As an example, we consider the sequence $x_k(t) = t^k$ for $0 \leq t \leq 1$ and $x_k(t) = 1$ for $1 \leq t \leq 2$. Then $\{x_k\}_{k=1}^{\infty}$ is a Cauchy sequence in $C[0,2]$ but does not converge in $C[0,2]$ with respect to $(\cdot,\cdot)_{L^2}$ since, with respect to the L^2-norm, it converges to the function

$$x(t) \;=\; \begin{cases} 0, \, t < 1, \\ 1, \, t \geq 1, \end{cases}$$

which is not continuous. The space $\big(C[a,b], \|\cdot\|_\infty\big)$, however, is a Banach space.

Every normed space or pre-Hilbert space X can be "completed," i.e., there exists a "smallest" Banach or Hilbert space $\tilde{X}$, respectively, that extends X (i.e., $\|x\|_X = \|x\|_{\tilde{X}}$ or $(x,y)_X = (x,y)_{\tilde{X}}$, respectively, for all $x, y \in X$). More precisely, we have the following result.

Theorem A.7. *Let X be a normed space with norm $\|\cdot\|_X$. There exists a Banach space $(\tilde{X}, \|\cdot\|_{\tilde{X}})$ and an injective (i.e. one-to-one) linear operator $j : X \to \tilde{X}$ such that*

(i) The range $j(X) \subset \tilde{X}$ is dense in $\tilde{X}$, and
(ii) $\|jx\|_{\tilde{X}} = \|x\|_X$ for all $x \in X$, i.e., j preserves the norm.

Furthermore, $\tilde{X}$ is uniquely determined in the sense that if $\hat{X}$ is a second space with properties (i) and (ii) with respect to a linear operator $\hat{j}$, then the operator $\hat{j}\, j^{-1} : j(X) \longrightarrow \hat{j}(X)$ has an extension to a norm-preserving isomorphism from $\tilde{X}$ onto $\hat{X}$. In other words, $\tilde{X}$ and $\hat{X}$ can be identified.

If X is a pre-Hilbert space then $\tilde{X}$ is a Hilbert space, and the operator j is unitary in the sense that $(jx, jy)_{\tilde{X}} = (x, y)_X$ for all $x, y \in X$.

Example A.8. As an example, we consider the completion of the pre-Hilbert space $C[a, b]$ with respect to the norm of the inner product

$$(x, y)_{L^2(a,b)} = \int_a^b x(t)\, \overline{y(t)}\, dt, \quad x, y \in C[a, b].$$

This completion is denoted by $L^2(a, b)$. We note that this definition of $L^2(a, b)$ is of purely functional analytic character. The advantage of this definition is obvious: The space $L^2(a, b)$ is complete (i.e. a Hilbert space) and contains $C[a, b]$ as a dense subspace. An equivalent and more direct approach uses the Lebesgue integration theory and will be sketched in the next section.

Definition A.9. *(Separable Space)The normed space X is called* **separable** *if there exists a countable dense subset $M \subset X$, i.e., if there exist M and a bijective mapping $j : \mathbb{N} \to M$ with $\overline{M} = X$.*

The spaces $\mathbb{C}^n$, $\mathbb{R}^n$, $L^2(a, b)$, and $C[a, b]$ are all separable. For the first two examples, let M consist of all vectors with rational coefficients; for the latter examples, take polynomials with rational coefficients.

Definition A.10. *(Orthogonal Complement)Let X be a pre-Hilbert space (over $\mathbb{F} = \mathbb{R}$ or $\mathbb{C}$).*

(a) Two elements x and y are called **orthogonal** *if $(x, y) = 0$.*
(b) Let $M \subset X$ be a subset. The set

$$M^\perp := \{x \in X : (x, y) = 0 \text{ for all } y \in M\}$$

is called the **orthogonal complement** *of M.*

$M^\perp$ is always a closed subspace and $M \subset (M^\perp)^\perp$. Furthermore, $A \subset B$ implies that $B^\perp \subset A^\perp$.

The following theorem is a fundamental result in Hilbert space theory and relies heavily on the completeness property.

Theorem A.11. *(Projection Theorem)*

Let X be a Hilbert space and $V \subset X$ be a closed subspace. Then $V = \left(V^\perp\right)^\perp$. Every $x \in X$ possesses a unique decomposition of the form $x = v + w$, where $v \in V$ and $w \in V^\perp$. The operator $P : X \to V$, $x \mapsto v$, is called the **orthogonal projection operator** *onto V and has the properties*

(a) $Pv = v$ for $v \in V$, i.e., $P^2 = P$;
(b) $\|x - Px\| \le \|x - v\|$ for all $v \in V$.

This means that $Px \in V$ is the best approximation of $x \in X$ in the closed subspace V.

There exists a generalization of this result to best approximations in convex sets. First, we recall the definition of a convex set.

Definition A.12. *A subset $U \subset X$ of a vector space X is called* **convex** *if*

$$\lambda x + (1 - \lambda) y \in U \quad \textit{for all } x, y \in U \textit{ and } \lambda \in [0, 1]\,.$$

Theorem A.13. *Let X be a Hilbert space and $U \subset X$ be a closed convex set. Then, for every $x \in X$ there exists a unique $\hat{u} \in U$ such that $\|x - \hat{u}\| \le \|x - u\|$ for all $u \in U$. Furthermore, $\hat{u} \in U$ is characterized by the variational inequality*

$$\operatorname{Re}\left(\hat{u} - x\,,\,\hat{u} - u\right) \;\le\; 0 \quad \textit{for all } u \in U\,.$$

Proof: Note that $\iota := \inf_{u \in U} \|u - x\| \ge 0$ so that the function $u \to \|u - x\|$ is bounded below on U. Let $u^{(1)}, u^{(2)}, \ldots$ be a sequence of points of U such that $\lim_{i \to \infty} \|x - u^{(i)}\| = \iota$. Then, by the parallelogram equality,

$$\begin{aligned} \|u^{(i)} - u^{(j)}\|^2 &= \left\|(u^{(i)} - x) - (u^{(j)} - x)\right\|^2 \\ &= 2\,\|u^{(i)} - x\|^2 + 2\,\|u^{(j)} - x\|^2 - 4\left\|\frac{1}{2}\left(u^{(i)} + u^{(j)}\right) - x\right\|^2 . \end{aligned}$$

Since U is convex, $\frac{1}{2}\left(u^{(i)} + u^{(j)}\right) \in U$ so that $\left\|\frac{1}{2}(u^{(i)} + u^{(j)}) - x\right\| \ge \iota$. Hence

$$\|u^{(i)} - u^{(j)}\|^2 \;\le\; 2\,\|u^{(i)} - x\|^2 + 2\,\|u^{(j)} - x\|^2 - 4\,\iota^2\,.$$

As $i, j \to \infty$, we have $2\,\|u^{(i)} - x\|^2 + 2\,\|u^{(j)} - x\|^2 - 4\,\iota^2 \to 0$. Thus, $\{u^{(j)}\}_{j=1}^\infty$ is a Cauchy sequence and has a limit point u. Since U is closed, $u \in U$. Moreover, since the function $u \to \|u - x\|$ is a continuous function from $H \to \mathbb{R}$,

$$\iota \;=\; \lim_{j \to \infty} \|u^{(j)} - x\| \;=\; \|u - x\|\,.$$

In order to show uniqueness of the point with minimal norm, suppose that there were two points, $u, v \in U$, $u \ne v$, such that $\|u - x\| = \|v - x\| = \iota$. Then, again by the parallelogram equality,

$$\begin{aligned} 0 < \|u-v\|^2 &= 2\,\|u-x\|^2 + 2\,\|v-x\|^2 - 4\left\|\frac{1}{2}(u+v)-x\right\|^2 \\ &= 2\,\iota^2 + 2\,\iota^2 - 4\left\|\frac{1}{2}(u+v)-x\right\|^2 , \end{aligned}$$

so that $4\,\iota^2 > 4\left\|\frac{1}{2}(u+v)-x\right\|^2$ or $\left\|\frac{1}{2}(u+v)-x\right\| < \iota$ which would give a vector in U whose distance to x is less than the infimum ι. □

A.3 The Lebesgue Integral and Function Spaces

A.3.1 The Lebesgue Integral

There are several ways to introduce the idea of the Lebesgue integral and the primary reason for doing so in applied analysis is that its behavior with respect to limiting operations is significantly better than that of the classical Riemann integral. We cannot pretend to give a treatment of the Lebesgue integral here. Rather, we give just some basic definitions and statements of results which illustrate this advantage and which explain the use of the integral in the text. We do not prove any of the results but refer to, e.g. the monograph [142].

We will confine our discussion to the real line, $\mathbb{R}$, and start with a simple definition.

Definition A.14. *A subset $N \subset \mathbb{R}$ is said to have* **measure zero** *provided that, for any $\epsilon > 0$ the set N can be covered by a finite or countably infinite set of intervals whose total length does not exceed ϵ. We call these sets also* **zero sets**.

In what follows, the behavior of functions on sets of measure zero are ignored, as that behavior is not important with respect to the integral. A simple example of such a set is a set containing finite or countably many points of $\mathbb{R}$.

Now, given a finite (or infinite) interval $(a,b) \subset \mathbb{R}$, we consider the class of **step functions** φ which are piecewise constant functions, i.e. for which finitely many subintervals $I_k \subset (a,b)$, $k = 1,\dots,p$, and constants c_k, $k = 1,\dots,p$, exist with $I_j \cap I_k = \emptyset$ for $j \neq k$ and $(a,b) = \bigcup_{k=1}^{p} I_k$ and

$$\varphi(t) := \begin{cases} c_k\,, & t \in I_k,\ k = 1,\dots,p, \\ 0, & \text{otherwise.} \end{cases}$$

Obviously, the Lebesgue integral of f is defined by

$$\int\limits_a^b \varphi(t)\,dt = \sum_{k=1}^{p} c_k\,|I_k|$$

From this definition it is clear that it does not matter to which of the subintervals the discontinuities belong.

One first looks at pointwise convergence of functions.

Definition A.15. *Let $\{\varphi_k\}_{k=1}^{\infty}$ be a sequence of functions.*

(a) This sequence is said to be **monotonically decreasing,** *if $\varphi_{k+1}(t) \leq \varphi_k(t)$ for all $k \in \mathbb{N}$ and all $t \in (a,b) \setminus N$, where N is a set of measure zero.*
(b) This sequence is said to **converge almost everywhere** *in (a,b) to the function φ provided that $\varphi_k(t) \to \varphi(t)$ for all $t \in (a,b) \setminus N$, where N is a set of measure zero.*
(c) A function $f : (a,b) \to \mathbb{R}$ is called (Lebesgue-) **measurable** *if there exists a sequence of step functions which converge to f almost everywhere.*

It can be shown that sums, differences, products, and scalar multiples of measurable functions are again measurable. For step functions we have two very simple results (the proofs of which are, however, not simple at all):

Proposition A.16. *Let $\{\varphi_k\}_{k=1}^{\infty}$ be a sequence of step functions.*

(a) If this sequence converges monotonically to 0 almost everywhere, the sequence of values of the corresponding integrals converges to 0.
(b) If this sequence is increasing and has integrals which are bounded by a single bound, then the sequence of step functions converges almost everywhere to a finite limit f.

From this last proposition, it makes sense to *define* the **Lebesgue integral** of such a limiting function in terms of the limits of the integrals. Indeed, since the integrals of the step functions all have a common bound and since they are increasing, these numbers converge to a finite limit and so we make the definition:

$$\int_a^b f(t)\,dt \;:=\; \lim_{k\to\infty} \int_a^b \varphi_k(t)\,dt\,.$$

In order for this definition to make sense, we must of course check that the value of the integral is *independent* of the choice of sequence of step functions converging almost everywhere to the function f. It is possible, with little effort, to verify this fact.

We note that, from this definition, it is immediate that the integral is *additive* in the sense that

$$\int_a^b [f(t) + g(t)]\,dt \;=\; \int_a^b f(t)\,dt \;+\; \int_a^b g(t)\,dt\,.$$

This construction effectively extends the definition of the integral to the class of functions which are limits of increasing sequences of step functions. The next step is to extend the integral to functions which can be written as *differences* of functions of this latter class. Given two such functions, f_1 and f_2, we define

$$\int_a^b [f_1(t) - f_2(t)]\, dt := \int_a^b f_1(t)\, dt - \int_a^b f_2(t)\, dt .$$

From the additivity of the integral defined before, it is easy to see that if g_1 and g_2 are such that $f_1(t) - f_2(t) = g_1(t) - g_2(t)$ almost everywhere, then

$$\int_a^b [f_1(t) - f_2(t)]\, dt = \int_a^b [g_1(t) - g_2(t)]\, dt .$$

The functions for which his integral are defined are called **integrable** (in the sense of Lebesgue) or **summable**.

A complex-valued functions $f : (a,b) \to \mathbb{C}$ is said to be integrable if its real- and imaginary parts are integrable and we obviously define

$$\int_a^b f(t)\, dt = \int_a^b \operatorname{Re} f(t)\, dt + i \int_a^b \operatorname{Im} f(t)\, dt .$$

The basic properties of the Lebesgue integral are given next.

Theorem A.17. *The set of (real or complex valued) integrable functions is a vector space over $\mathbb{R}$ or $\mathbb{C}$, respectively. Moreover, if f, defined on (a,c) is integrable over (a,b) and also integrable over (b,c) for some $b \in (a,c)$ then it is integrable over the interval (a,c) and*

$$\int_a^c f(t)\, dt = \int_a^b f(t)\, dt + \int_b^c f(t)\, dt .$$

Furthermore, it is **absolutely integrable** *in the sense that, if f is integrable, then so is $|f|$ and the* **triangle inequality** *holds*

$$\left| \int_a^c f(t)\, dt \right| \leq \int_a^c |f(t)|\, dt .$$

Finally, for every integrable function, there exists a sequence of step functions $\{\varphi_k\}_{k=1}^{\infty}$ such that $\varphi_k(t) \to f(t)$ almost everywhere and

$$\int_a^b |f(t) - \varphi_k(t)|\, dt \longrightarrow 0 \quad \textit{as } k \longrightarrow \infty .$$

One can show that, if a function is integrable in the sense of Riemann, then it is integrable in the sense of Lebesgue and that the values of the integrals coincide. What is particularly important about the Lebesgue integral is its behavior with respect to convergence of functions. There are two main results which deal with sequences of integrable functions (not just step functions).

Theorem A.18. *(Beppo-Levi)*
Every increasing sequence $\{f_n\}_{n=1}^{\infty}$ of integrable functions on the interval (a,b) whose integrals have a common bound, converges almost everywhere to an integrable function f and

$$\int_a^b f(t)\,dt \;=\; \lim_{n\to\infty} \int_a^b f_n(t)\,dt\,.$$

It is a corollary of this theorem that $\int_a^b |f(t)|\,dt = 0$ if and only if $f(t) = 0$ almost everywhere in (a,b).

Finally, we have the most often used convergence result which we have used several times in the text. Note that there is no *a priori* assumption that the limit function in this theorem is integrable.

Theorem A.19. *(Lebesgue Dominated Convergence Theorem)*
Let $\{f_n\}_{n=1}^{\infty}$ be a sequence of integrable functions on the interval (a,b) which converge almost everywhere to a function f. Suppose further than there exists an integrable function g such that for almost all $t \in (a,b)$,

$$|f_n(t)| \;\leq\; g(t) \quad \textit{for all } n\,,$$

then the function f is integrable and

$$\int_a^b f(t)\,dt \;=\; \lim_{n\to\infty} \int_a^b f_n(t)\,dt\,.$$

A.3.2 Sobolev Spaces

Using the Lebesgue integration theory we can now define the space $L^2(a,b)$ as follows. First, we define the vector space

$$\mathcal{L}^2(a,b) \;:=\; \bigl\{x : (a,b) \to \mathbb{C} : x \text{ is measurable and } |x|^2 \text{ integrable}\bigr\},$$

Then $\mathcal{L}^2(a,b)$ is a vector space since, for $x,y \in \mathcal{L}^2(a,b)$ and $\alpha \in \mathbb{C}$, $x+y$ and αx are also measurable and αx, $x+y \in \mathcal{L}^2(a,b)$, the latter by the binomial theorem $|x(t)+y(t)|^2 \leq 2\,|x(t)|^2 + 2\,|y(t)|^2$. We define a sesquilinear form on $\mathcal{L}^2(a,b)$ by

$$\langle x,y\rangle \;:=\; \int_a^b x(t)\,\overline{y(t)}\,dt, \quad x,y\in\mathcal{L}^2(a,b).$$

$\langle\cdot,\cdot\rangle$ is not an inner product on $\mathcal{L}^2(a,b)$ since $\langle x,x\rangle = 0$ only implies that x vanishes almost everywhere, i.e., that $x\in\mathcal{N}$, where $\mathcal{N}$ is defined by

$$\mathcal{N} \;:=\; \bigl\{x\in\mathcal{L}^2(a,b) : x(t)=0 \text{ a.e. on } (a,b)\bigr\}.$$

Now we define $L^2(a,b)$ as the factor space

$$L^2(a,b) \;:=\; \mathcal{L}^2(a,b)/\mathcal{N}$$

and equip $L^2(a,b)$ with the inner product

$$\bigl([x],[y]\bigr)_{L^2} \;:=\; \int_a^b x(t)\,\overline{y(t)}\,dt, \quad x\in[x],\ y\in[y].$$

Here, $[x],[y]\in L^2(a,b)$ are equivalence classes of functions in $\mathcal{L}^2(a,b)$, i.e. $[x]=\bigl\{z\in\mathcal{L}^2(a,b) : z-x\in\mathcal{N}\bigr\}$. Then it can be shown that this definition is well-defined and yields an inner product on $L^2(a,b)$. From now on, we will write $x\in L^2(a,b)$ instead of $x\in[x]\in L^2(a,b)$. Furthermore, it can be shown that $L^2(a,b)$, defined in this way, is complete and contains $C[a,b]$ as a dense subspace. Therefore, it is the completion of $C[a,b]$ with respect to the inner product $(\cdot,\cdot)$ as it was defined in Example A.8.

In an analogous way, the spaces $L^p(a,b)$ for $p\geq 1$ can be defined by the two equivalent ways indicated above. In particular, it is the completion of $C[a,b]$ with respect to the norm

$$\|x\|_{L^p} \;=\; \left(\int_a^b |x(t)|^p\,dt\right)^{1/p}.$$

The space $L^\infty(a,b)$, however, has to be defined by the (Lebesgue) integration theory as follows: As we did it for $L^2(a,b)$, we define first the space $\mathcal{M}(a,b)$ of all functions, defined on the interval (a,b), taking values in $\mathbb{C}$ and measurable with respect to Lebesgue measure such that the **essential supremum** of x is finite. The essential supremum of $|x|$ is defined by

$$\operatorname{ess\,sup}|x| \;:=\; \inf\bigl\{m : \bigl|x(t)\bigr|\leq m \text{ a.e. on } (a,b)\bigr\}. \tag{A.7a}$$

Thus the essential supremum is the least number m such that the inequality $\bigl|x(t)\bigr|\leq m$ holds except on a set of measure zero.

Again, we set $\mathcal{N} := \bigl\{x\in\mathcal{M}(a,b) : x(t)=0 \text{ a.e. on } (a,b)\bigr\}$ and define the space $L^\infty(a,b)$ as the set of equivalence classes of functions of $\mathcal{M}(a,b)$, i.e. $L^\infty(a,b)=\mathcal{M}(a,b)/\mathcal{N}$. We equip $L^\infty(a,b)$ with norm

$$\|[x]\|_\infty \; := \; \operatorname{ess\,sup}|x| \quad \text{for any } x \in [x] . \tag{A.7b}$$

Again, we identify the equivalence class $[x]$ with x. With this definition, it can be shown that $L^\infty(a,b)$ is a Banach space which contains $C[a,b]$ as a closed proper subspace.

We now turn to the definition of **Sobolev spaces**. Again, there are different equivalent definitions. Instead of using distributional derivatives we again define them as the completion of spaces of smooth functions with respect to an integral norm using Theorem A.7:

Definition A.20. *Let $d \in \mathbb{N}$, and $\Omega \subset \mathbb{R}^d$ an open and bounded set. For every multi-index $q = (q_1, \dots, q_d) \in \mathbb{N}_0^d$ we define the* **differential operator**

$$D^q \; := \; \frac{\partial^{q_1+\cdots+q_d}}{\partial x_1^{q_1} \cdots \partial x_d^{q_d}} .$$

(a) For every $m \in \mathbb{N}$ the space $C^m(\overline{\Omega})$ is defined as the space of $m-$times in Ω continuously differentiable functions $x : \Omega \longrightarrow \mathbb{C}$ such that D^q can be extended to uniformly continuous functions on $\overline{\Omega}$ for every multi-index $q \in \mathbb{N}^d$ with $q_1 + \cdots + q_d \leq m$.

(b) For every $m \in \mathbb{N}$ the **Sobolev space** *$H^m(\Omega)$ is defined as the completion of $C^m(\overline{\Omega})$ with respect to norm, induced by the inner product*

$$(x,y)_{H^m(\Omega)} \; = \; (x,y)_{L^2(\Omega)} + \sum_{|q|\leq m} (D^q x, D^q y)_{L^2(\Omega)} , \quad x,y \in C^m(\overline{\Omega}) , \tag{A.8}$$

where $|q| = q_1 + \cdots + q_d$ and the L^2-inner product is defined as

$$(x,y)_{L^2(\Omega)} \; = \; \iint_\Omega x(t)\,\overline{y(t)}\,dt .$$

In the one-dimensional case, when $\Omega = (a,b)$ is an interval, the induced norm

$$\|x\|_{H^m(a,b)} \; = \; \left(\sum_{q=0}^{m} \|x^{(q)}\|^2_{L^2(a,b)} \right)^{1/2}$$

is equivalent to the norm

$$\|x\| \; = \; \left(\sum_{q=0}^{m-1} |x^{(q)}(a)|^2 \; + \; \|x^{(m)}\|_{L^2(a,b)} \right)^{1/2} .$$

Therefore, the Sobolev space $H^m(a,b)$ can also be characterized as

$$H^p(a,b) \; = \; \left\{ x \in C^{(p-1)}[a,b] : \begin{array}{c} x^{(p-1)}(t) = x^{(p-1)}(a) + \int_a^t z(s)\,ds , \\ a \leq t \leq b , \text{ for some } z \in L^2(a,b) \end{array} \right\} . \tag{A.9}$$

z is the generalized derivative of x of order m for which one write, of course, $x^{(m)}$. **Rellich's imbedding theorem** states that the Sobolev space $H^m(\Omega)$ is *compactly* imbedded in the Sobolev space $H^n(\Omega)$ whenever $n < m$ and the domain Ω is "smooth enough"[1]

We also define the Sobolev space $H_0^1(\Omega)$ of generalized functions with zero-boundary conditions as the completion of $\{x \in C^1(\overline{\Omega}) : x = 0 \text{ on } \partial\Omega\}$ with respect to the H^1−norm.

It can be shown that the **trace operator** $\gamma_0 : C(\overline{\Omega}) \longrightarrow C(\partial\Omega)$ which maps a continuous function x on $\overline{\Omega}$ onto the restriction on the boundary $\partial\Omega$ has a bounded extension from $H^1(\Omega)$ into $L^2(\partial\Omega)$. The range of this extension $\gamma_0 : H^1(\Omega) \longrightarrow L^2(\partial\Omega)$ is denoted by $H^{1/2}(\partial\Omega)$ and is equipped with the norm

$$\|x\|_{H^{1/2}(\partial\Omega)} = \inf\big\{\|\tilde{x}\|_{H^1(\Omega)} : \gamma_0\tilde{x} = x\big\}, \quad x \in H^{1/2}(\partial\Omega).$$

It's **dual space** is denoted by $H^{-1/2}(\partial\Omega)$. If $\partial\Omega$ is smooth enough, these space and, more generally, the spaces $H^s(\partial\Omega)$ for real values of s can be defined via local coordinates and the Fourier transform. The standard reference to Sobolev spaces is [3]. For an equivalent definition for smooth closed curves see also [74].

A.4 Orthonormal Systems

In this section, let X always be a *separable* Hilbert space over the field $\mathbb{F} = \mathbb{R}$ or $\mathbb{C}$.

Definition A.21. *(Orthonormal System)*

A countable set of elements $A = \{x_k : k \in \mathbb{N}\}$ *is called an* **orthonormal system** *if*

(i) $(x_k, x_j) = 0$ *for all* $k \neq j$ *and*
(ii) $\|x_k\| = 1$ *for all* $k \in \mathbb{N}$.

A is called a **complete** *or a* **maximal** *orthonormal system if, in addition, there is no orthonormal system,* B*, with* $A \subset B$ *and* $A \neq B$.

One can show using Zorn's Lemma (see [61]) that every separable Hilbert space possesses a maximal orthonormal system. Furthermore, it is well-known from linear algebra that every countable set of linearly independent elements of X can be ortho-normalized. The algorithm for doing so is called the Gram-Schmidt Orthogonalization Process.

For any set $A \subset X$, let

[1] This latter assumption holds, e.g., if Ω satisfies a cone condition (see [3]).

$$\operatorname{span} A \;:=\; \left\{ \sum_{k=1}^{n} \alpha_k \, x_k : \alpha_k \in \mathbb{F},\; x_k \in A,\; n \in \mathbb{N} \right\} \tag{A.10}$$

be the subspace of X spanned by A.

Theorem A.22. *Let $A = \{x_k : k = 1,2,3,\ldots\}$ be an orthonormal system. Then:*

(a) Every finite subset of A is linearly independent.
(b) If A is finite, i.e., $A = \{x_k : k = 1,2,\ldots,n\}$, then for every $x \in X$ there exist uniquely determined coefficients $\alpha_k \in \mathbb{F}$, $k = 1,\ldots,n$, such that

$$\left\| x - \sum_{k=1}^{n} \alpha_k x_k \right\| \;\le\; \|x - a\| \quad \textit{for all } a \in \operatorname{span} A. \tag{A.11}$$

The coefficients α_k are given by $\alpha_k = (x, x_k)$ for $k = 1,\ldots,n$ and are called the **(generalized) Fourier coefficients** *of x with respect to the orthonormal system A.*
(c) For every $x \in X$, the following **Bessel inequality** *holds:*

$$\sum_{k=1}^{\infty} |(x,x_k)|^2 \;\le\; \|x\|^2 . \tag{A.12}$$

In particular, the series converges in $\mathbb{R}$.
(d) A is complete if and only if $\operatorname{span} A$ is dense in X.
(e) A is complete if and only if for all $x \in X$ **Parseval's equation** *holds:*

$$\sum_{k=1}^{\infty} |(x,x_k)|^2 \;=\; \|x\|^2 . \tag{A.13}$$

(f) A is complete if and only if every $x \in X$ has a **(generalized) Fourier expansion** *of the form*

$$x \;=\; \sum_{k=1}^{\infty} (x, x_k)\, x_k , \tag{A.14}$$

where the convergence is understood in the norm of X. In this case, Parseval's equation holds in the following more general form:

$$(x,y) \;=\; \sum_{k=1}^{\infty} (x,x_k)\,\overline{(y,x_k)} . \tag{A.15}$$

This important theorem includes, as special examples, the classical Fourier expansion of periodic functions and the expansion with respect to orthogonal polynomials. We recall these two examples.

Example A.23. (Fourier Expansion)
(a) The functions $x_k(t) := \exp(ikt)/\sqrt{2\pi}$, $k \in \mathbb{Z}$, form a complete system of orthonormal functions in $L^2(0, 2\pi)$ over $\mathbb{C}$. By part (f) of the previous theorem, every function $x \in L^2(0, 2\pi)$ has an expansion of the form

$$x(t) \;=\; \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} \mathrm{e}^{ikt} \int_0^{2\pi} x(s)\, \mathrm{e}^{-iks} ds,$$

where the convergence is understood in the sense of L^2, i.e.,

$$\int_0^{2\pi} \left| x(t) \;-\; \frac{1}{2\pi} \sum_{k=-M}^{N} \mathrm{e}^{ikt} \int_0^{2\pi} x(s)\, \mathrm{e}^{-iks} ds \right|^2 dt \;\longrightarrow\; 0$$

as M, N tend to infinity. For smooth periodic functions uniform convergence can also be shown.

(b) The **Legendre polynomials** P_k, $k = 0, 1, \ldots$, form a maximal orthonormal system in $L^2(-1, 1)$. They are defined by

$$P_k(t) \;=\; \gamma_k \frac{d^k}{dt^k}(1 - t^2)^k, \quad t \in (-1, 1),\; k \in \mathbb{N} \cup \{0\},$$

with normalizing constants

$$\gamma_k \;=\; \sqrt{\frac{2k+1}{2}}\, \frac{1}{k!\, 2^k}\,.$$

We refer to [50] for details.

Other important examples will be given later.

A.5 Linear Bounded and Compact Operators

For this section, let X and Y always be normed spaces and $A : X \to Y$ be a linear operator.

Definition A.24. *(Boundedness, Norm of A)*
The linear operator A is called **bounded** *if there exists $c > 0$ such that*

$$\|Ax\| \;\leq\; c\,\|x\| \quad \textit{for all } x \in X\,.$$

The smallest of these constants is called the **norm** *of A, i.e.,*

$$\|A\| \;:=\; \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|}\,. \tag{A.16}$$

Theorem A.25. *The following assertions are equivalent:*

(a) A is bounded.
(b) A is continuous at $x = 0$, i.e., $x_j \to 0$ implies that $Ax_j \to 0$.
(c) A is continuous for every $x \in X$. .

The space $\mathcal{L}(X,Y)$ of all linear bounded mappings from X to Y with the operator norm is a normed space, i.e., the operator norm has the usual properties of a norm as well as the following: Let $B \in \mathcal{L}(X,Y)$ and $A \in \mathcal{L}(Y,Z)$, then $A\,B \in \mathcal{L}(X,Z)$ and $\|A\,B\| \leq \|A\|\,\|B\|$.

Integral operators are the most important examples for our purposes.

Theorem A.26. *(a) Let $k \in L^2\big((c,d) \times (a,b)\big)$. The operator*

$$(Ax)(t) \;:=\; \int\limits_a^b k(t,s)\,x(s)\,ds\,, \quad t \in (c,d)\,, \quad x \in L^2(a,b)\,, \tag{A.17}$$

is well-defined, linear, and bounded from $L^2(a,b)$ into $L^2(c,d)$. Furthermore,

$$\|A\|_{L^2} \;\leq\; \int\limits_c^d \int\limits_a^b |k(t,s)|\;ds\,dt\,.$$

(b) Let k be continuous on $[c,d] \times [a,b]$. Then A is also well-defined, linear, and bounded from $C[a,b]$ into $C[c,d]$ and

$$\|A\|_\infty \;=\; \max_{t\in[c,d]} \int\limits_a^b |k(t,s)|\;ds\,.$$

We can extend this theorem to integral operators with weakly singular kernels. We recall that a kernel k is called **weakly singular** on $[a,b] \times [a,b]$ if k is defined and continuous for all $t,s \in [a,b]$, $t \neq s$, and there exist constants $c > 0$ and $\alpha \in [0,1)$ such that

$$|k(t,s)| \;\leq\; c\,|t-s|^{-\alpha} \quad \text{for all } t,s \in [a,b],\ t \neq s.$$

Theorem A.27. *Let k be weakly singular on $[a,b] \times [a,b]$. Then the integral operator A, defined by (A.17) for $[c,d] = [a,b]$, is well-defined and bounded as an operator in $L^2(a,b)$ as well as in $C[a,b]$.*

The statement of this last theorem suggests that if we have a bounded linear operator defined on a normed linear space, it is possible to extend this operator to the completion of that space. Indeed, we can state the following theorem which makes this statement precise.

Theorem A.28. *Let X and Y be normed spaces with completions $\tilde{X}$ and $\tilde{Y}$ respectively, and suppose that $A \in \mathcal{L}(X,Y)$. Then there exists a unique extension $\tilde{A} \in \mathcal{L}(\tilde{X},\tilde{Y})$ with $\|\tilde{A}\| = \|A\|$.*

For the special case $Y = \mathbb{F}$, we denote by $X^* := \mathcal{L}(X,\mathbb{F})$ the **dual space** of X. Analogously, the space $X^{**} := \big(X^*\big)^*$ is called the **bidual** of X. The canonical embedding $j : X \to X^{**}$, defined by

$$(jx)x^* \;:=\; x^*(x)\,, \quad x \in X\,,\; x^* \in X^*\,,$$

is linear, bounded, one-to-one, and satisfies $\|jx\| = \|x\|$ for all $x \in X$. For the latter property, one needs the Hahn-Banach Theorem (in particular Corollary A.42 below).

We pause to give two examples of Banach spaces and their duals.

Examples A.29. (a) Let $p \in [1,\infty)$ and $\Gamma \in \mathbb{R}^d$ some measurable set. Then an important result from integration theory states that the dual of $L^p(\Gamma)$ is isomorphic to $L^q(\Gamma)$ where $q = \infty$ if $p = 1$, and q is determined by $1/p+1/q = 1$ if $p \in (1,\infty)$. The dual form is given by

$$(x^*,x) \;=\; \int\limits_{\Gamma} \overline{x^*(t)}\, x(t)\, dt\,, \quad x \in L^p(\Gamma)\,,\; x^* \in L^q(\Gamma)\,.$$

(b) We look at the example of the Banach space c_0 (see (A.4a)) over $\mathbb{R}$. Assume that $x^* \in c_0^*$ and let $\boldsymbol{e}^{(k)}$ be the usual unit vectors. Then, for any $x \in c_0$,

$$\lim_{n\to\infty} \sum_{k=1}^{n} x_k\, \boldsymbol{e}^{(k)} \;=\; x\,,$$

and by continuity of the functional x^*,

$$x^*(x) \;=\; \lim_{n\to\infty} x^*\left(\sum_{k=1}^{n} x_k\, \boldsymbol{e}^{(k)}\right) \;=\; \lim_{n\to\infty} \left(\sum_{k=1}^{n} x_k\, x^*(\boldsymbol{e}^{(k)})\right).$$

For simplicity, define $\xi_k := x^*(\boldsymbol{e}^{(k)})$, $k \in \mathbb{N}$, and define the sequence $\{x_k^{(N)}\}_{k=1}^{\infty} \subset c_0$ of elements in c_0 by

$$x_k^{(N)} \;:=\; \begin{cases} \dfrac{|\xi_k|}{\xi_k}\,, & \text{if } k \le N \text{ and } \xi_k \ne 0\,, \\[2ex] 0\,, & \text{if } k > N \text{ or } \xi_k = 0\,. \end{cases}$$

Then $\|x^{(N)}\| \le 1$ and $\|x^*\|_{c_0^*} = \sup\limits_{\|x\|\le 1} |x^*(x)| \ge \big|x^*(x^{(N)})\big| = \sum_{k=1}^{N} |\xi_k|$. Letting $N \to \infty$, it follows that $y := \{\xi_k\}_{k=1}^{\infty} \in \ell^1$ and that $\|y\|_{\ell^1} = \sum_{k=1}^{\infty} |\xi_k| \le \|x^*\|_{c_0^*}$.

Conversely, if $y = \{y_k\}_{k=1}^{\infty} \in \ell^1$, then for any $x = \{x_k\}_{k=1}^{\infty} \in c_0$,

$$\left| \sum_{k=1}^{\infty} x_k \, y_k \right| \;\leq\; \|x\|_{c_0} \, \|y\|_{\ell^1} \,,$$

and so y defines an element $x^* \in c_0^*$ and, moreover, $\|x^*\|_{c_0^*} \leq \|y\|_{\ell^1}$. This computation shows that ℓ^1 is the dual space of c_0.

Definition A.30. *(Reflexive Space)*

The normed space X is called **reflexive** *if the canonical embedding, j, is onto, i.e., a norm-preserving isomorphism from X onto the bidual space X^{**}.*

The following important result gives a characterization of X^* in Hilbert spaces.

Theorem A.31. *(Riesz-Fischer)*

Let X be a Hilbert space. For every $x \in X$, the functional $x^*(y) := (y, x)$, $y \in X$, defines a linear bounded mapping from X to $\mathbb{F}$, i.e., $x^* \in X^*$. Furthermore, for every $x^* \in X^*$ there exists one and only one $x \in X$ with $x^*(y) = (y, x)$ for all $y \in X$ and

$$\|x\| \;=\; \|x^*\| \;:=\; \sup_{y \neq 0} \frac{|x^*(y)|}{\|y\|} \,.$$

It is instructive to look at a concrete example.

Example A.32. Consider the complex Hilbert space $L^2(-\pi, \pi)$. Then, the Riesz-Fischer theorem says that, for a given $x^* \in \bigl(L^2(-\pi, \pi)\bigr)^*$, there is a unique $x \in L^2(-\pi, \pi)$ such that, for all $y \in L^2(-\pi, \pi)$, we have

$$x^*(y) \;=\; \int_{-\pi}^{\pi} y(t) \, \overline{x(t)} \, dt \,.$$

The Riesz-Fischer Theorem implies that every Hilbert space is reflexive. It also yields the existence of a unique adjoint operator for every linear bounded operator $A : X \longrightarrow Y$.

Theorem A.33. *(Adjoint Operator)*

Let $A : X \longrightarrow Y$ be a linear and bounded operator between Hilbert spaces. Then there exists one and only one linear bounded operator $A^* : Y \longrightarrow X$ with the property

$$(Ax, y)_Y \;=\; (x, A^*y)_X \quad \textit{for all } x \in X, \; y \in Y \,.$$

This operator $A^* : Y \longrightarrow X$ is called the **adjoint operator** *to A. For $X = Y$, the operator A is called* **self-adjoint** *if $A^* = A$.*

Example A.34. (a) Let $X = L^2(a,b)$, $Y = L^2(c,d)$, and $k \in L^2\big((c,d) \times (a,b)\big)$. The adjoint A^* of the integral operator

$$(Ax)(t) \;=\; \int_a^b k(t,s)\, x(s)\, ds\,, \quad t \in (c,d)\,, \quad x \in L^2(a,b)\,,$$

is given by

$$(A^*y)(t) \;=\; \int_c^d \overline{k(s,t)}\, y(s)\, ds\,, \quad t \in (a,b)\,, \quad y \in L^2(c,d)\,.$$

(b) Let the space $X = C[a,b]$ of continuous function over $\mathbb{C}$ be supplied with the L^2-inner product. Define $f, g : C[a,b] \to \mathbb{R}$ by

$$f(x) \;:=\; \int_a^b x(t)\, dt \quad \text{and} \quad g(x) := x(a) \quad \text{for } x \in C[a,b]\,.$$

Both f and g are linear. However f is bounded while g is unbounded. According to Theorem A.28 there is an extension of f to a bounded linear functional (also denoted by f) on $L^2(a,b)$, i.e., $f \in \big(L^2(a,b)\big)^*$. By Theorem A.33, we can identify $\big(L^2(a,b)\big)^*$ with $L^2(a,b)$ itself. For the given f, the representation function is just the constant function 1 since $f(x) = (x,1)_{L^2}$ for $x \in L^2(a,b)$. The adjoint of f is calculated by

$$f(x) \cdot \overline{y} \;=\; \int_a^b x(t)\, \overline{y}\, dt \;=\; (x,y)_{L^2} \;=\; \big(x, f^*(y)\big)_{L^2}$$

for all $x \in L^2(a,b)$ and $y \in \mathbb{C}$. Therefore, $f^*(y) \in L^2(a,b)$ is the constant function with value y.

(c) Let X be the **Sobolev space** $H^1(a,b)$, i.e., the space of L^2-functions that possess generalized L^2-derivatives:

$$H^1(a,b) \;:=\; \left\{ x \in L^2(a,b) : \begin{array}{l} \text{there exists } \alpha \in \mathbb{F} \text{ and } y \in L^2(a,b) \text{ with} \\ x(t) = \alpha + \int_a^t y(s)\, ds \text{ for } t \in (a,b) \end{array} \right\}.$$

We denote the generalized derivative $y \in L^2(a,b)$ by x'. We observe that $H^1(a,b) \subset C[a,b]$ with bounded embedding. As an inner product in $H^1(a,b)$, we define

$$\big(x,y\big)_{H^1} \;:=\; x(a)\,\overline{y(a)} \;+\; (x',y')_{L^2}, \quad x, y \in H^1(a,b)\,.$$

Now let $Y = L^2(a,b)$ and $A : H^1(a,b) \longrightarrow L^2(a,b)$ be the operator $x \mapsto x'$ for $x \in H^1(a,b)$. Then A is well-defined, linear, and bounded. It is easily seen that the adjoint of A is given by

$$(A^*y)(t) = \int_a^t y(s)\,ds\,, \quad t \in (a,b)\,, \quad y \in L^2(a,b)\,.$$

For the remaining part of this section, we will assume that X and Y are normed spaces and $K : X \to Y$ a linear and bounded operator.

Definition A.35. *(Compact Operator)*

The operator $K : X \to Y$ is called **compact** *if it maps every bounded set S into a relatively compact set $K(S)$.*

We recall that a set $M \subset Y$ is called **relatively compact** if every **bounded** sequence $\{y_j\}_{j=1}^{\infty} \subset M$ has an accumulation point in $\overline{M}$, i.e., if the closure $\overline{M}$ is compact.

From the definition we see that a compact operator is automatically bounded. The converse statement is, however, true if and only if the space X is finite dimensional.

Theorem A.36.

(a) If X is finite dimensional, and Y a normed space then every linear operator $K : X \to Y$ is compact.
(b) Let X be a normed space. Then the identity operator $I : X \to X$ is compact if and only if X is finite dimensional.

The set of all compact operators from X into Y is a closed subspace of $\mathcal{L}(X,Y)$ and is even invariant with respect to composition with a bounded linear operator in the sense given in the following theorem.

Theorem A.37.

(a) If K_1 and K_2 are compact operators from X into Y, then so are $K_1 + K_2$ and λK_1 for every $\lambda \in \mathbb{F}$.
(b) Let $K_n : X \longrightarrow Y$ be a sequence of compact operators between Banach spaces X and Y. Let $K : X \longrightarrow Y$ be bounded, and let K_n converge to K in the **operator norm**, *i.e.,*

$$\|K_n - K\| = \sup_{x \neq 0} \frac{\|K_n x - Kx\|}{\|x\|} \longrightarrow 0\,, \quad \text{as } n \to \infty\,.$$

Then K is also compact.
(c) If $L \in \mathcal{L}(X,Y)$ and $K \in \mathcal{L}(Y,Z)$, and L or K is compact, then KL is also compact.
(d) Let $A_n \in \mathcal{L}(X,Y)$ be **pointwise convergent** *to some $A \in \mathcal{L}(X,Y)$, i.e., $A_n x \to Ax$ for all $x \in X$. If $K : Z \to X$ is compact, then $\|A_n K - A\,K\| \to 0$, i.e., the operators $A_n K$ converge to AK in the operator norm.*

The integral operators give important examples of compact operators.

Theorem A.38.

(a) Let $k \in L^2\big((c,d) \times (a,b)\big)$. The operator $K : L^2(a,b) \to L^2(c,d)$, defined by

$$(Kx)(t) \;:=\; \int_a^b k(t,s)\,x(s)\,ds\,, \quad t \in (c,d)\,, \quad x \in L^2(a,b)\,, \tag{A.18}$$

is compact from $L^2(a,b)$ into $L^2(c,d)$.

(b) Let k be continuous on $[c,d] \times [a,b]$ or weakly singular on $[a,b] \times [a,b]$ (in this case $[c,d] = [a,b]$). Then K defined by (A.18) is also compact as an operator from $C[a,b]$ into $C[c,d]$.

We will now study equations of the form

$$x \;-\; Kx \;=\; y\,, \tag{A.19}$$

where the linear operator $K : X \to X$ is either small or compact. The first theorem is sometimes referred to as the **perturbation theorem** and is related to the **Neumann series**.

Theorem A.39. *Let X be a Banach space and $K : X \to X$ be a linear operator such $\|K\| < 1$. Then the limit*

$$S \;:=\; \lim_{n\to\infty} \sum_{j=0}^{n} K^n \tag{A.20a}$$

exists in the norm of $\mathcal{L}(X,X)$ and $S \in \mathcal{L}(X,X)$ is the inverse of $I - K$. Furthermore,

$$\|(I-K)^{-1}\| \;\leq\; \frac{1}{1 - \|K\|}\,. \tag{A.20b}$$

The following theorem extends the well-known existence results for finite linear systems of n equations and n variables to compact perturbations of the identity.

Theorem A.40. *(Riesz)*

Let X be a normed space and $K : X \to X$ be a linear compact operator.

(a) The nullspace $\mathcal{N}(I-K) = \big\{x \in X : x = Kx\big\}$ is finite-dimensional and the range $(I-K)(X)$ is closed in X.

(b) If $I - K$ is one-to-one, then $I - K$ is also surjective, and the inverse $(I - K)^{-1}$ is bounded. In other words, if the homogeneous equation $x - Kx = 0$ admits only the trivial solution $x = 0$, then the inhomogeneous equation $x - Kx = y$ is uniquely solvable for every $y \in X$ and the solution x depends continuously on y.

A.6 The Hahn-Banach Theorem

The Hahn-Banach Theorem has been called the most important theorem of Functional Analysis for the study of optimization problems. The theorem can be stated in two apparently different forms, the *extension form* and the *geometric form.* Each of these forms has several important corollaries so that in a certain sense, when one refers to the Hahn-Banach Theorem, one refers to an entire set of results and uses the particular version appropriate to the particular application. Indeed, we have used both forms in this book. The theorem is true in very general settings, far more general than the versions stated here. But these versions are those which are particularly useful to us.

The extension form of the Hahn-Banach Theorem is concerned with the extension of a bounded linear functional, defined on a proper subspace V of a normed space X, to a linear functional defined on the entire space without increasing the norm of the linear functional. The norm of an element $x \in V$ is the same as its norm with respect to the entire Banach space, i.e., $\|x\|_V = \|x\|_X$. Moreover, if we have a bounded linear functional defined on such a subspace V, then $x^* \in V^*$ and

$$\|x^*\|_{V^*} := \sup_{y \in V} \frac{|x^*(y)|}{\|y\|_X} .$$

The precise statement of the extension result is the following.

Theorem A.41. *(Hahn-Banach)*

Let X be a normed linear space over $\mathbb{F} = \mathbb{R}$ or $\mathbb{C}$, and let x^* be a bounded linear functional defined on a closed subspace V of X with $V \neq X$. Then x^* has an extension $\hat{x}^* \in X^*$, i.e., $\hat{x}^*(x) = x^*(x)$ for all $x \in V$, which satisfies $\|x^*\|_{V^*} = \|\hat{x}^*\|_{X^*}$.

The first important fact to follow from this theorem is that there are non-zero elements of X^* and, in fact, sufficiently many to separate points.

Corollary A.42.
Let X be a Banach space over $\mathbb{F}$.

(a) For every $x \in X$, $x \neq 0$, there is an $x^* \in X^*$, $x^* \neq 0$, such that $x^*(x) = \|x^*\|_{X^*} \, \|x\|_X$.
(b) For any $x, y \in X$, $x \neq y$, there exists a functional $x^* \in X^*$ such that $x^*(x) \neq x^*(y)$.

The first claim of the corollary is easily established if we consider the one-dimensional subspace $V = \text{span}\{x\}$, and define $x^*(\alpha x) := \alpha \, \|x\|$. Then $\|x^*\|_{V^*} = 1$ and so x^* has an extension to all of X of unit norm.

The second part follows easily from the first by looking at the non-zero element $x - y \in X$. □

In order for us to set the stage for the geometric form of the Hahn-Banach Theorem, we need the definition of a hyperplane.

Definition A.43. *Let X be a normed space and $V \subset X$ a subspace.*

(a) Any set of the form $x_o + V := \{x_o + v : v \in V\}$ with $x_o \in X$ is called a **linear manifold** *in X.*
(b) Any linear manifold $x_o + V$ is called a **hyperplane** *provided there is no subspace $W \subset X$ with $V \subsetneqq W \subsetneqq X$.*
(c) In the case that V is a closed subspace, we refer to the linear manifolds as **closed** *linear manifolds.*

Example A.44. If $X = \mathbb{R}^n$ considered as a real vector space, $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ a matrix where $m \leq n$, and $\boldsymbol{b} \in \mathbb{R}^m$ then $M = \{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}\}$ is a closed linear manifold. In $\mathbb{R}^2$ this can be a point (if $m = n = 2$ and $\boldsymbol{A}$ regular) or a line or all of $\mathbb{R}^2$. For general $n = 3$ and $m = 1$ the linear manifold M reduces to $M = \{\boldsymbol{x} \in \mathbb{R}^3 : \boldsymbol{a}^\top \boldsymbol{x} = \boldsymbol{b}\}$ where $\boldsymbol{a} \in \mathbb{R}^3$. If $\boldsymbol{a} \neq \boldsymbol{n}$ then M describes a plane in $\mathbb{R}^3$ in normal form

From this simple example, we can see that there is an intimate connection between the geometric notion of a closed hyperplane and that of a non-zero bounded linear functional on X. Indeed, we have a correspondence between the two described in the next proposition.

Proposition A.45.
If $x^* \in X^*$ and $x^* \neq 0$, and $c \in \mathbb{F}$ then the set $H := \{x \in X : x^*(x) = c\}$ is a closed hyperplane in X. Conversely, if H is a closed hyperplane in X, then there exist a linear functional $x^* \in X^*$ and a constant $c \in \mathbb{F}$ such that $H = \{x \in X : x^*(x) = c\}$.

For any closed hyperplane, H, we can normalize c so that $c \in \mathbb{R}$. Even more, if $0 \notin H$, then $c \neq 0$ and we can normalize further by taking $\hat{x}^* = x^*/c$ so that the equation defining the closed hyperplane is $\hat{x}^*(x) = 1$.

In the case that the normed space X is over $\mathbb{F} = \mathbb{R}$, i.e., is a *real* normed linear space, the hyperplane H determines open and closed **half-spaces**, namely:

$$H_{\leq c} := \{x \in X : x^*(x) \leq c\}, \quad \text{and} \quad H_{<c} := \{x \in X : x^*(x) < c\}, \tag{A.21a}$$

and

$$H_{\geq c} := \{x \in X : x^*(x) \geq c\}, \quad \text{and} \quad H_{>c} := \{x \in X : x^*(x) > c\}. \tag{A.21b}$$

The use of the term half-space is justified by the observation that

$$\{x \in X : x^*(x) \leq c\} \cup \{x \in X : x^*(x) \geq c\} = X,$$

and

$$\{x \in X : x^*(x) \leq c\} \,\cap\, \{x \in X : x^*(x) \geq c\} \;=\; H\,.$$

We remark that of the half-spaces defined in (A.21), the half- spaces $H_{\leq c}$ and $H_{\geq c}$ are closed while the others are open. This observation follows immediately from the fact that the functional x^* is continuous.

With these ideas in hand, we can state the basic geometric form of the Hahn-Banach theorem which is also called the **strict separation theorem**.

Theorem A.46. *(Strict Separation Theorem)*
Let X be a normed space over $\mathbb{F}$, $K \subset X$ be a closed convex set, and $x_o \in X \setminus K$. Then there exists a closed hyperplane $H := \big\{x \in X : x^*(x) = c\big\}$ with $x^* \neq 0$ and $c \in \mathbb{R}$ such that

$$\operatorname{Re}\big[x^*(x_0)\big] \;<\; c \;\leq\; \operatorname{Re}\big[x^*(x)\big] \quad \textit{for all } x \in K\,.$$

Remark: If K is a closed subspace we can replace x by λx in this inequality (for any $\lambda \in \mathbb{F}$) from which it easily follows that even $x^*(x) = 0$ for all $x \in K$. In this case $c = 0$ can be chosen.

Example A.47.
As an example we consider the solvability of a linear system to determine $x \in X$ such that

$$x_k^*(x) \;=\; \alpha_k\,, \quad k = 1, 2, \dots, n\,, \tag{A.22}$$

where $x_k^* \in X^*$, and $\alpha_k \in \mathbb{C}$ are given. Define a map $T : X \to \mathbb{C}^n$ by

$$T(x) \;:=\; \big(x_1^*(x)\,, \dots, x_n^*(x)\big)^\top\,, \quad x \in X\,.$$

Let $\mathcal{R} = T(X)$ denote the range of T and note that $\mathcal{R}$ is a finite-dimensional subspace of $\mathbb{C}^n$ of dimension n and therefore closed.

We claim that the system is solvable, i.e. $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_n)^\top \in \mathcal{R}$, if and only if, the following condition (C) holds:

For every vector $(\lambda_1, \lambda_2, \dots, \lambda_n)^\top \in \mathbb{C}^n$ with $\sum_{k=1}^n \lambda_k\, x_k^* = 0$ we have that $\sum_{k=1}^n \lambda_k\, \alpha_k = 0$.

Indeed, suppose that this condition (C) holds but the system (A.22) is not solvable. Then $\alpha \notin \mathcal{R}$. It follows from the Hahn-Banach Theorem A.46 that there exists an element $\hat{y}^* \in \mathbb{C}^n$ such that $\hat{y}^*(\alpha) < 0$ and $\hat{y}^*(y) = 0$ for all $y \in \mathcal{R}$. Since $\hat{y}^*$ can be represented by an vector of the form $(\lambda_1, \lambda_2, \dots, \lambda_n)^\top \in \mathbb{C}^n$, this means that

$$\hat{y}(y) \;=\; \sum_{k=1}^{\infty} \lambda_k\, y_k \;=\; 0 \text{ for all } y \in \mathcal{R} \quad \text{and} \quad \hat{y}(\alpha) \;=\; \sum_{k=1}^{\infty} \lambda_k\, \alpha_k \;<\; 0\,,$$

which contradicts the condition.

Conversely, suppose that the system (A.22) is solvable, and that for all $\lambda \in \mathbb{C}^n$ we have that $\sum_{k=1}^n \lambda_k \, x_k^* = 0$. Then $\alpha \in \mathcal{R}$ so that there exists an $x \in X$ such that $T(x) = \alpha$, i.e. $x_k^*(x) = \alpha_k$ for all $k = 1, \dots, n$. Therefore,

$$0 \;=\; \sum_{k=1}^{\infty} \lambda_k \, x_k^*(x) \;=\; \sum_{k=1}^{\infty} \lambda_k \, \alpha_k \,,$$

which shows that the condition (C) is satisfied.

A.7 The Fréchet Derivative

In this section, we will briefly recall some of the most important results for nonlinear mappings between normed spaces. The notions of continuity and differentiability carry over in a very natural way.

Definition A.48. *Let X and Y be normed spaces over the field $\mathbb{F} = \mathbb{R}$ or $\mathbb{C}$, $U \subset X$ an open subset, $\hat{x} \in U$, and $T : X \supset U \to Y$ be a (possibly nonlinear) mapping.*

(a) T is called **continuous** *in $\hat{x}$ if for every $\epsilon > 0$ there exists $\delta > 0$ such that $\|T(x) - T(\hat{x})\| \leq \epsilon$ for all $x \in U$ with $\|x - \hat{x}\| \leq \delta$.*

(b) T is called **Fréchet differentiable** *at $\hat{x}$ if there exists a linear bounded operator $A : X \to Y$ (depending on $\hat{x}$) such that*

$$\lim_{h \to 0} \frac{1}{\|h\|} \big\|T(\hat{x} + h) - T(\hat{x}) - Ah\big\| \;=\; 0\,. \tag{A.23}$$

We write $T'(\hat{x}) := A$. In particular, $T'(\hat{x}) \in \mathcal{L}(X, Y)$.

(c) The mapping T is called **continuously Fréchet differentiable** *for $\hat{x} \in U$ if T is Fréchet differentiable in a neighborhood V of $\hat{x}$ and the mapping $T' : V \to \mathcal{L}(X, Y)$ is continuous in $\hat{x}$.*

Continuity and differentiability of a mapping depend on the norms in X and Y, in contrast to the finite-dimensional case. If T is differentiable in $\hat{x}$, then the linear bounded mapping A in part (b) of Definition A.48 is unique. Therefore, $T'(\hat{x}) := A$ is well-defined. If T is differentiable in x, then T is also continuous in x. In the finite-dimensional case $X = \mathbb{F}^n$ and $Y = \mathbb{F}^m$, the linear bounded mapping $T'(x)$ is given by the Jacobian (with respect to the Cartesian coordinates).

We should remark that it is often the case in the applications as those we have considered in this book, we are confronted with the situation that the map $T : X \to Y$ maps a *complex* Banach space X into a *real* Banach space Y. In particular, it is often the case that $Y = \mathbb{R}$ considered as a vector space over itself. Difficulties arise in this situation because of the definition of linearity of the operator $T'(\hat{x}) : X \to Y$. In this situation, one considers the space X as a real vector space by restricting the field of scalars. If we denote this space by $X_{\mathbb{R}}$, then there exists a linear map $T'(\hat{x}) : X_{\mathbb{R}} \to Y$, with

$$\frac{\|T(\hat{x}+h) - T(\hat{x}) - T'(\hat{x})\,h\|}{\|h\|} \longrightarrow 0\,, \quad \text{as } \|h\| \to 0\,.$$

Example A.49.

(a) Let $f : [c,d] \times [a,b] \times \mathbb{C} \to \mathbb{C}$ be continuous and continuously differentiable with respect to the third argument. Let the mapping $T : C[a,b] \to C[c,d]$ be defined by

$$T(x)(t) \;:=\; \int\limits_a^b f\big(t,s,x(s)\big)\,ds\,, \quad t \in [c,d]\,,\; x \in C[a,b]\,.$$

Then T is continuously Fréchet differentiable with derivative

$$\big(T'(x)z\big)(t) \;=\; \int\limits_a^b \frac{\partial}{\partial x} f\big(t,s,x(s)\big)\,z(s)\,ds\,, \quad t \in [c,d]\,,\; x,z \in C[a,b]\,.$$

(b) Let X and Y be Hilbert spaces over $\mathbb{F}$ and let $K : X \to Y$ be a bounded linear operator and fix $y \in Y$. Define $T : X \to \mathbb{R}$ by

$$T(x) \;=\; \|Kx - y\|^2\,, \quad x \in X\,.$$

Then

$$T'(x)\,h \;=\; 2\,\mathrm{Re}\,(Kx - y, Kh) \;=\; 2\,\mathrm{Re}\,\big(K^*(Kx-y), h\big)\,, \quad x,h \in X\,.$$

Indeed, by the Binomial Theorem

$$\begin{aligned} T(\hat{x}+h) - T(\hat{x}) &= \|K(\hat{x}+h) - y\|^2 - \|K\hat{x} - y\|^2 \\ &= 2\,\mathrm{Re}\,(K\hat{x} - y, Kh) + \|h\|^2 \\ &= 2\,\mathrm{Re}\,\big(K^*\,(K\hat{x} - y), h\big) \;+\; \|h\|^2\,, \end{aligned}$$

and the map $T'(\hat{x}) : X_{\mathbb{R}} \to \mathbb{R}$ defined by $h \mapsto 2\,\mathrm{Re}\,\big(K^*\,(K\hat{x}-y), h\big)$ is linear.

The following theorem collects further properties of the Fréchet derivative.

Theorem A.50.

(a) Let $T, S : X \supset U \to Y$ be Fréchet differentiable for $x \in U$. Then $T + S$ and λT are also Fréchet differentiable for all $\lambda \in \mathbb{F}$ and

$$(T+S)'(x) \;=\; T'(x) + S'(x)\,, \qquad (\lambda T)'(x) \;=\; \lambda T'(x)\,.$$

(b) **Chain rule**: *Let $T : X \supset U \to V \subset Y$ and $S : Y \supset V \to Z$ be Fréchet differentiable for $x \in U$ and $T(x) \in V$, respectively. Then ST is also Fréchet differentiable in x and*

$$(ST)'(x) \;=\; \underbrace{S'\big(T(x)\big)}_{\in \mathcal{L}(Y,Z)}\ \underbrace{T'(x)}_{\in \mathcal{L}(X,Y)} \;\in\; \mathcal{L}(X,Z)\,.$$

(c) *Special case: If $T : X \to Y$ is Fréchet differentiable for $\hat{x} \in X$, then so is $\psi : \mathbb{F} \to Y$, defined by $\psi(t) := T(t\hat{x})$, $t \in \mathbb{F}$, for every point $t \in \mathbb{F}$ and $\psi'(t) = T'\big(t\hat{x}\big)\hat{x} \in Y$. Note that originally $\psi'(t) \in \mathcal{L}(\mathbb{F}, Y)$. In this case, one identifies the linear mapping $\psi'(t) : \mathbb{F} \to Y$ with its generating element $\psi'(t) \in Y$.*

A.8 Weak Convergence

The familiar theorem that a continuous functional on a compact set achieves both its maximum and minimum values is only useful if there are enough sets which are compact. Unfortunately, the usual definition of (sequential) compactness which involves norm convergence is only valid in finite dimensional spaces. The requirement that a set be weakly sequentially compact or weak*-sequentially compact are much less restrictive, and therefore useful in many infinite dimensional contexts.

Given a normed space X and its dual space X^*, a sequence $\{x_n\}_{n=1}^{\infty} \subset X$ is said to converge **strongly**, or **with respect to the norm**, to the point $x \in X$ provided $\|x - x_n\| \to 0$ as $n \to \infty$ (see Definition A.2. For optimization problems, the weak convergence is important.

Definition A.51. *Let X be a normed space.*

(a) *A sequence $\{x_n\}_{n=1}^{\infty} \subset X$ is said to converge* **weakly** *(or converge in the weak topology) to $x \in X$ provided $x^*(x_n) \to x^*(x)$ for all $x^* \in X^*$. Weak convergence is often denoted by $x_n \rightharpoonup x$.*
(b) *A set $U \subset X$ is called* **weakly sequentially closed**, *if the limit point of every weakly convergent sequence $\{x_n\}_{n=1}^{\infty} \subset U$ belongs to U.*
(c) *A set $U \subset X$ is called* **weakly sequentially compact**, *if every sequence $\{x_n\}_{n=1}^{\infty} \subset U$ contains a weakly convergent subsequence such that its limit point belongs to U.*
(d) *A functional $\mathcal{J} : X \to \mathbb{R}$ is called* **weakly sequentially continuous** *provided for every sequence $\{\psi_k\}_{k=1}^{\infty}$ converging weakly to an element $\psi \in X$ we have*

$$\lim_{k\to\infty} \mathcal{J}(\psi_k) \;=\; \mathcal{J}(\psi)\,.$$

Using the characterization of the dual space, H^*, of a Hilbert space H given by the theorem of Riesz-Fischer (Theorem A.31), we can express weak convergence in terms of the inner product. In this context, using the inner product on the Hilbert space, we can write:

$$(x_n, x^*) \longrightarrow (x, x^*) \quad \text{for all } x^* \in H\,,$$

where $(\cdot,\cdot)$ is the given inner product on H.

It is clear that any strongly convergent sequence is weakly convergent since we have the obvious inequality

$$|x^*(x) - x^*(x_n)| \;\leq\; \|x^*\|_{X^*}\, \|x - x_n\|_X\,.$$

Example A.52. Consider the Hilbert space ℓ^2 of sequences $\boldsymbol{x} = (x_1, x_2, \dots)$, $x_k \in \mathbb{C}$, with $\sum_{k=1}^{\infty} |x_k|^2 < \infty$. The standard basis vectors $\boldsymbol{e}^{(n)} := (0, 0, \dots, 0, 1, 0, \dots)$, $n = 1, 2, \dots$, where the only non-zero entry, 1, occurs in the n^{th} position, lie in the unit ball $B[0,1]$ of ℓ^2. However there is no convergent subsequence since $\|\boldsymbol{e}^{(m)} - \boldsymbol{e}^{(n)}\| = \sqrt{2}$ for $m \neq n$. In particular, the sequence itself does not converge. Now, choose *any* $\boldsymbol{x} \in \ell^2$. Then, we know by the definition that $\sum_{k=1}^{\infty} |x_k|^2 < \infty$, so that $x_k \to 0$ as $k \to \infty$. For any k, the inner product $(\boldsymbol{e}^{(k)}, \boldsymbol{x}) = x_k$ and so, for any $\boldsymbol{x} \in \ell^2$, we have $(\boldsymbol{e}^{(n)}, \boldsymbol{x}) \to 0$ as $n \to \infty$. Hence $\boldsymbol{e}^{(n)} \rightharpoonup 0$ weakly in ℓ^2.

Remark: This example presents a definite contrast to the situation for $\mathbb{R}^n$ and $\mathbb{C}^n$. There, the unit ball is closed and bounded and such sets are compact according to the Bolzano-Weierstrass Theorem. In fact, compactness of the unit ball *characterizes* finite dimensional spaces as we have seen in Theorem A.5. Also, we note that in finite dimensional spaces weak and norm convergence are the same. Indeed, if $\boldsymbol{x}^{(k)} \rightharpoonup \boldsymbol{x}$ in $\mathbb{C}^n$, then $x_m^{(k)} = \bigl(\boldsymbol{x}^{(k)}, \boldsymbol{e}^{(m)}\bigr) \to \bigl(\boldsymbol{x}, \boldsymbol{e}^{(m)}\bigr) = x_m$ for every $m = 1, 2, \dots, n$. Therefore, all components of $\boldsymbol{x}$ converge in $\mathbb{C}$ which implies norm convergence of $\{\boldsymbol{x}^{(k)}\}_{k=1}^{\infty}$ to $\boldsymbol{x}$ in any norm on $\mathbb{C}^n$.

Example A.52 makes it clear that weak convergence of a sequence does not imply that the sequence converges in norm. On the other hand, given a weakly convergent sequence in a Banach space, we can construct a *new* sequence of points, using the elements of the original one, which converges in norm to the weak limit. The result is due to Mazur.

Theorem A.53. *Let X be a real Banach space and $\{x_n\}_{n=1}^{\infty} \subset X$ a sequence which converges weakly to $x \in X$. Then there exists a system of real numbers $\alpha_{k,i} \geq 0$, $i = 1, 2, \dots, k$, $k = 1, 2, \dots$, with $\sum_{i=1}^{k} \alpha_{k,i} = 1$, such that,*

$$y_k \;=\; \sum_{i=1}^{k} \alpha_{k,i}\, x_i\,, \qquad \textit{and } y_k \to x\,,$$

or equivalently, $\|y_k - x\| \to 0$, as $k \to \infty$.

Besides the notion of weak convergence, sometimes the idea of weak*-convergence is also useful in optimization problems.

In order to understand the situation in Banach or Hilbert space, we recall the construction of the bidual X^{**}. We may, of course, consider the weak convergence in X^* as we did with the pair $\{X, X^*\}$, but it turns out to be more fruitful to consider a type of convergence defined, not by *all* the functionals in X^{**}, but only by the functionals in that space generated by the elements $x \in X \subset X^{**}$. This is possible since the canonical embedding $j : X \to X^{**}$ has a range, $j(X) \subset X^{**}$ which is an isometric copy of X so that we may consider $X \subset X^{**}$. In other words, we consider only functionals in X^{**}, of the form $x^{**}(x^*) := x^*(x)$, $x^* \in X^*$, for each $x \in X$. Weak convergence in X^* with respect to these particular functions is called weak*-convergence. Precisely,

Definition A.54. *A sequence of elements $\{x_n^*\}_{n=1}^{\infty} \subset X^*$ is said to converge in the weak* sense to $x^* \in X^*$ provided, $x_n^*(x) \to x^*(x)$, for every $x \in X$. The definitions of weak*-sequential closedness or weak*-sequential compactness of a set is defined just as in Definition A.51.*

Example A.55. Consider the Banach space, c_0, of all infinite sequences of real numbers $\{x_k\}_{k=1}^{\infty}$ such that $x_k \to 0$ as $k \to \infty$, and with norm $\|\boldsymbol{x}\| := \max_{k \in \mathbb{N}} |x_k|$. The dual space $c_0^* = \ell^1$ as we have seen in Example A.29. One can show, by arguments similar to those in that example that the dual space of ℓ^1 is ℓ^∞. So the bidual $c_0^{**} = (\ell^1)^* = \ell^\infty$. Let $\boldsymbol{e}^{(n)}$ be the usual sequence with 1 in the n^{th} entry and zeros elsewhere, and take $\boldsymbol{x}_n^* = \boldsymbol{e}^{(n)}$. Then $\boldsymbol{x}_n^* \to 0$ in the weak*-sense of Definition A.54, but $\boldsymbol{x}_n^*$ does *not* converge weakly to 0 since, for $\boldsymbol{x}^{**} = (1,1,1,\ldots,1,\ldots) \in \ell^\infty$ we have $\boldsymbol{x}^{**}(\boldsymbol{x}_n^*) = \big(\boldsymbol{e}^{(n)}, \boldsymbol{x}^{**}\big) = 1$ for all $n \in \mathbb{N}$.

The fact that the unit ball in the dual of a Banach space is compact with respect to weak*-convergence is due to Alaoglu (see [145]). We formulate it in the following way.

Theorem A.56. *Let X be a Banach space.*[2] *Then every bounded sequence $\{x_k^*\} \subset X^*$ in X^* contains a weak*-convergent subsequence.*

We apply this result to bounded sequences in $L^p(\Gamma)$ for $p \in (1, \infty]$ (including $p = \infty$) by noting that $L^p(\Gamma)$ is the dual of $L^q(\Gamma)$ with $1/p + 1/q = 1$ (where $p = \infty$ belongs to $q = 1$). The dual form is given by the extension of the L^2–inner product

$$(x^*, x) \;=\; \int\limits_{\Gamma} \overline{x^*(t)}\, x(t)\, dt\,, \quad x \in L^q(\Gamma)\,,\ x^* \in L^p(\Gamma)\,,$$

see Example A.29.

[2] Note that we always assume that the spaces are separable.

Corollary A.57. *Let $\Gamma \subset \mathbb{R}^d$ be some Lebesgue measurable set, $p \in (1,\infty]$ and $q \in [1,\infty)$ such that $1/p + 1/q = 1$. Then every bounded sequence $\{x_k^*\}_{k=1}^{\infty} \subset L^p(\Gamma)$ contains a subsequence $\{x_{k_j}^*\}_{j=1}^{\infty}$ and some $x^* \in L^p(\Gamma)$ such that*

$$\int_{\Gamma} \overline{x_{k_j}^*(t)}\, x(t)\, dt \longrightarrow \int_{\Gamma} \overline{x^*(t)}\, x(t)\, dt\,,\ j \to \infty\,,$$

for all $x \in L^q(\Gamma)$.

What is particularly important for our work is the following result for a reflexive Banach space (and hence for a Hilbert space). We recall that a Banach space is reflexive if it is isometrically isomorphic to its bidual under the canonical embedding.

Theorem A.58. *Let X be a (separable) reflexive Banach space or, in particular, a Hilbert space. Then the unit ball $B[0,1] \subset X$ is weakly sequentially compact.*

Remark: This result follows immediately from Theorem A.56 and the definition of reflexivity which says that, under the canonical imbedding, X and X^{**} are isometrically isomorphic.

A.9 Partial Orderings

This section contains the basic definitions and properties needed for the material in Chapter 8. The exposition is somewhat formal in the interest of efficiency, but we give illustrations of the main points throughout in order to aid the reader's understanding.

We present the results in the setting of a real or complex vector (linear) space. It is possible to develop the theory in a significantly more general setting, but there is not real need here for such generality. On the other hand, the reader may always be more concrete and replace the general vector space with $\mathbb{R}^n$.

Whatever setting is chosen, the motivation is always the same; we are interested in being able to compare the efficiency of different choices of inputs (or strategies) when there are several performance indices which need to be considered. In the *scalar* case i.e., when there is only one real-valued performance index, we work in the vector space $\mathbb{R}$ which has a natural order, and the comparison between two choices of strategy is simple. In the more general case, we must leave this familiar ground.

In general, it is not possible to impose a concrete total ordering[3] and we must content ourselves with a so-called *partial ordering.*

[3] There is a statement, due to Zermelo, that any set can be well-ordered. This statement has the nature of an existence theorem and gives no practical way to *produce* such an ordering. Indeed the statement is equivalent to the famous Axiom of Choice. (See [61]).

To be more precise, we start with the definition of a general relation on an arbitrary set.

Definition A.59. *Let S be any set and $\mathcal{P}$ a relation defined on S i.e., $\mathcal{P}$ is a subset of the set of all ordered* pairs *of elements $(s_1, s_2) \in S \times S$. Then we say that s_1* **precedes** *s_2 if the ordered pair $(s_1, s_2) \in \mathcal{P}$. We will use the notation $s_1 \prec s_2$ if $(s_1, s_2) \in \mathcal{P}$.*

Given such a relation (and we have introduced the word "precedes" advisedly) it may have certain properties. In particular, it may have the properties that make it suitable for us to use as a (partial) ordering.

Definition A.60. *Given a set S and a relation $\prec$ defined as above, we say that $\prec$ defines a* **partial ordering** *of the set S provided:*

(i) for all $s \in S$, $s \prec s$ (reflexivity),
(ii) if $s_1, s_2 \in S$ and both $s_1 \prec s_2$ and $s_2 \prec s_1$ then $s_1 = s_2$ (anti-symmetry),
(iii) if $s_1, s_2, s_3 \in S$ and if $s_1 \prec s_2$ and $s_2 \prec s_3$, then $s_1 \prec s_3$ (transitivity).

Otherwise said, a relation $\prec$ is a partial ordering on S if it is a reflexive, anti-symmetric, and transitive relation defined on S.

We should point out that the usual ordering on $\mathbb{R}$ is a partial ordering of that set. It is in fact *more* than that since any two elements are related, one being less than the other. In this case we have a **total-** or **well-ordering**. The difference is that, with a partial ordering, not every two elements are necessarily comparable.

The notion of a partial ordering is independent of the nature of the set S. However, in our applications we work in a vector space, Z, and we want to find a systematic way to impose a partial ordering on it. We first introduce the notion of a *cone*. This definition is dependent only on the algebraic structure of the vector space.

Definition A.61. *Let Z be a* **real** *linear space, and let $\Lambda \subset Z$ be non-empty. Then Λ is called a* **cone with vertex** *$0 \in Z$ provided that for every $z \in \Lambda$, and $\lambda > 0$ we have $\lambda z \in \Lambda$.*

We will find that it is useful to use the standard notation $-\Lambda := \{-z : z \in \Lambda\}$.

It is easily seen that the set $\{0\} \subset Z$ satisfies the definition of a cone. On the other hand, the entire vector space Z is *also* a cone and that instance is trivial as well. We are obviously interested in less trivial cases, for example the set $\Lambda \subset \mathbb{R}$ defined by

$$\Lambda := \mathbb{R}_{\geq 0}, \tag{A.24}$$

which is clearly a non-trivial cone in $\mathbb{R}$. Indeed, it generates the usual ordering in $\mathbb{R}$ as will become clear below.

More important for our work is the generalization of (A.24) given by

$$\Lambda \;:=\; \mathbb{R}^n_{\geq 0} \;:=\; \left\{\boldsymbol{x} \in \mathbb{R}^n : x_j \geq 0\,,\; j = 1, 2, \ldots, n\right\} \qquad \text{(A.25)}$$

which we will refer to as the **usual order cone** in $\mathbb{R}^n$. The cone Λ is just the first quadrant in the case that $n = 2$.

We isolate two properties that a cone may have since each one, separately, relates to one of the defining properties of a partial order.

Definition A.62. *A cone $\Lambda \subset Z$ with vertex 0 is called* **non-trivial** *provided $\Lambda \neq \{0\}$ and $\Lambda \neq Z$. The cone is called* **line-free** *provided $0 \in \Lambda$ and $\Lambda \cap (-\Lambda) = \{0\}$.*

It is easy to see that the cone described above in (A.25), namely $\Lambda := \left\{x \in \mathbb{R}^n : x_j \leq 0\,,\; j = 1, 2, \ldots, n\right\}$ is non-trivial and is line-free. Indeed, one need only notice that $-\Lambda$ is just $\left\{x \in \mathbb{R}^n : x_j \leq 0\,,\; j = 1, 2, \ldots, n\right\}$.

Example A.63. The set $\hat{\Lambda} := \mathbb{R}_{>0}$ is also a cone, is still convex, but does not contain the origin. In $\mathbb{R}^2$ the set

$$\left\{\boldsymbol{x} \in \mathbb{R}^2 : x_1 \geq 0\right\} \;\cup\; \left\{\boldsymbol{x} \in \mathbb{R}^2 : x_1 \leq 0,\; x_2 \geq -x_1\right\}$$

is an example of a cone that is not convex. Moreover, it fails to be line-free since it contains a line, namely the line $x_1 = 0$.

The point of introducing these definitions is to show that a cone, with these properties, can be used to define a partial ordering in Z. Indeed, given a cone Λ with vertex 0, we may define a binary relation $\prec$ by

$$x \prec y \quad \text{provided} \quad y - x \in \Lambda\,. \qquad \text{(A.26)}$$

With this definition $\prec$ we can easily check that this binary relation is a partial ordering of the vector space Z provided Λ is convex, contains the origin, and is line free.

(a) If $0 \in \Lambda$, then $\prec$ is reflexive. This follows from the observation that for any $x \in Z$, $x - x = 0 \in \Lambda$ which implies that $x \prec x$.

(b) If Λ is convex then $\prec$ is transitive, for if $x, y, z \in Z$, and if $x \prec y$ and $y \prec z$ then $y - x \in \Lambda$ and $z - y \in \Lambda$. Since Λ is convex,

$$\frac{1}{2}(y - x) \;+\; \frac{1}{2}(z - y) \;\in\; \Lambda\,,$$

and so $\frac{1}{2}(z - x) \in \Lambda$ from which it follows that $z - x \in \Lambda$. Hence $x \prec z$.

(c) If Λ is line-free, then $\prec$ is antisymmetric. Indeed, if $x \prec y$ and $y \prec x$ then $y - x \in \Lambda \cap (-\Lambda) = \{0\}$ so that $x = y$.

To summarize, these three observations show that the following theorem is true.

Theorem A.64. *If Z is a linear space and $\Lambda \subset Z$ is a line-free, convex cone with $0 \in \Lambda$, then the binary relation $\prec$ defined by*

$$x \prec y \quad \textit{if and only if} \quad y - x \in \Lambda\,,$$

defines a partial order on the vector space Z.

There is also the partial converse of this theorem. If $\prec$ is a partial order on X which respects the operations. i.e. $x \prec y \Rightarrow x + z \prec y + z$ and $\lambda x \prec \lambda y$ for all $x, y, z \in X$ and $\lambda > 0$ then $\Lambda := \{x \in X : 0 \prec x\}$ is a line-free, convex cone, and contains 0. This is easily proven by arguments similar to those above.

This leads to the following standard terminology:

Definition A.65. *A pair $\{Z, \prec\}$ where Z is a real linear space and $\prec$ is a partial order defined on Z is called an* **ordered vector space**. *The cone which induces the partial order is called the* **order cone**. *We write $<_\Lambda$ in place of $\prec$ to emphasize the relationship.*

We conclude this section by giving some further examples.

Example A.66. Let $Z = L^2(0,1)$, and choose

$$\Lambda \;=\; \big\{x \in L^2(0,1) : x(t) \geq 0 \text{ almost everywhere in } [0,1]\big\}\,.$$

Then Λ is a cone which contains 0, is line-free, and is convex. Indeed, it is easy to convince oneself that Λ is a line-free cone. However to check that it is convex some care is needed. Let $x, y \in L^2(0,1)$ and let $N(x)$ and $N(y)$ be the two sets of measure zero where x and y respectively do not satisfy the pointwise equality of the definition of the cone. Then $N(x) \cup N(y)$ is also of measure zero so that, for any $\lambda \in [0,1]$, $z := (1-\lambda)x + \lambda y$ is non-negative on $[0,1] \setminus [N(x) \cup N(y)]$. Hence, $z \in \Lambda$.

Notice that, with this order cone, the functions $x(t) = t$ and $y(t) = t^2$ are comparable and $y <_\Lambda x$. On the other hand, the functions $x(t) = t$ and $y(t) = \cos(\pi\, t)$ are not comparable with respect to $<_\Lambda$.

Example A.67. Let $Z = SL_n(\mathbb{R}^n)$, the set of symmetric $n \times n$ real matrices, and let $\Lambda = \{\boldsymbol{A} \in SL_n(\mathbb{R}^n) : \boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} \geq 0 \text{ for all } \boldsymbol{x} \in \mathbb{R}^n\}$ be the set of all positive semidefinite matrices. Then Λ is a convex, line-free cone and $0 \in \Lambda$ as the reader can check.

References

1. M. Abramowitz, and I. A. Stegun, *Handbook of Mathematical Functions*, Dover, 1970.
2. N.I. Achieser, *Theory of Approximation*, Ungar, 1956.
3. R. A. Adams, *Sobolev Spaces*, Academic Press, New York, 1975.
4. T. S. Angell, X. Jiang, and R. E. Kleinman, On a numerical method for inverse acoustic scattering. *Inverse Problems* **13** (1997), 531–545.
5. T. S. Angell, and A. Kirsch, The conductive boundary condition for Maxwell's equations. *SIAM J. Appl. Math.* **52** (1992), 1597–1610.
6. T. S. Angell, and A. Kirsch, Multicriteria optimization in antenna problems. *Math. Meth. in the Appl. Sciences* **15** (1992), 647–660.
7. T. S. Angell, and A. Kirsch, Optimization of radiating fields and the example of null-placement. In: *Analytical and Computational Methods in Scattering and Applied Mathematics*, F. Santosa and I. Stakgold, eds., Chapman and Hall/CRC Press, Boca Raton, FL, 2000.
8. T. S. Angell, and R. E. Kleinman, Generalized exterior boundary–value problems and optimization for the Helmholtz equation. *J. Optim. Theory Appl.* **37** (1982), 469–497.
9. T. S. Angell, and R. E. Kleinman, The Helmholtz equation with L_2 – –boundary values. *SIAM J. Math. Anal.* **16** (1985), 259–278.
10. T. S. Angell, and M. Z. Nashed, Operator theoretic and computational aspects of ill-posed problems in antenna theory. In: *International Symposium of Mathematical Theory of Networks and Systems* **3** P. Dewilde, ed., Western Periodicals Co., Los Angeles, 1979.
11. T. S. Angell, R. E. Kleinman, and G. F. Roach, An inverse transmission problem for the Helmholtz equation, *Inverse Problems* **3** (1987), 149–180.
12. T. S. Angell, A. Kirsch, and R. E. Kleinman, Antenna control and optimization. *Proc. IEEE* **79** (1991), 1559–1568.
13. T. S. Angell, R. E. Kleinman, and A. Kirsch, Multicriteria optimization in arrays. *Proceedings Journées Internationales de Nice sur les Antennes*, Nice, France 1992.
14. J. P. Aubin, and A. Cellina, *Differential Inclusions; Set–Valued Maps and Viability Theory*, Springer–Verlag, Berlin, Heidelberg, New York, Tokyo, 1986.
15. K. Azis, and R. B. Kellogg, Finite element analysis of a scattering problem. *Math. Comp.* **37** (1981), 261–272.

16. C. A. Balanis, *Antenna Theory: Analysis and Design, 2nd. ed.*, John Wiley & Sons, Inc., New York, 1997.
17. A. Baylis, C. Goldstein, and E. Turkel, Preconditioned conjugate gradient methods for the Helmholtz equation. In: *Elliptic Problem Solvers II*, G. Birkhoff and A. Schoenstadt, eds., Orlando, 1984, Academic Press, 233–243.
18. C. J. Bouwkamp, and N. G. de Bruijn, The problem of optimum current distribution. *Philips Res. Rep.* **1** (1945-46), 135–158.
19. J. H. Bramble, The Lagrange multiplier method for Dirichlet's problem. *Math. Comp.* **37** (1981), 1–11.
20. J. H. Bramble, and J. E. Pasciak, A new computational approach for the linearized scalar potential formulation of the magnetostatic field problem. *IEEE Trans. on Magnetics* **MAG–18** (1982), 357–361.
21. A. Calderon, Multiple expansion of radiation fields. *J. Rat. Mech. Anal.* **3** (1954), 523–537.
22. L. Cesari, *Optimization–Theory and Applications*, Springer–Verlag, Berlin, Heidelberg, New York, 1983.
23. M. Cessenat, *Mathematical Methods in Electromagnetism: Linear Theory and Applications*, World Scientific, Singapore, 1996.
24. P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, vol. 4 of Studies in Mathematics and It's Applications, Elsevier North–Holland, New York, 1978.
25. L. Collatz, Approximation in partial differential equations. In: *On Numerical Approximation*, R. E. Langer, ed., Madison, 1959, University of Wisconsin Press, 413–422.
26. L. Collatz, *The Numerical Treatment of Differential Equations*, Springer–Verlag, Berlin, Heidelberg, New–York, 1960.
27. R. E. Collin, and F. J. Zucker, *Antenna Theory: Part I*, McGraw–Hill Book Company, New York, St. Louis, San Francisco, 1969
28. F. Collino, and P. Monk, The perfectly matched layer in curvilinear coordinates, *Technical Report 3049*, INRIA, 1996.
29. D. Colton, and R. Kress, *Integral Equation Methods in Scattering Theory*, John Wiley & Sons, New York, 1983.
30. D. Colton, and R. Kress, *Inverse Acoustic and Electromagnetic Scattering*, Springer–Verlag, Berlin, Heidelberg, New York, 2nd. edition, 1998.
31. J. P. Dauer, and W. Stadler, A survey of vector optimization in infinite dimensional spaces, II. *J. Optim. Theory Appl.* **51** (1986), 205–241.
32. Ph. J. Davis, and P. Rabinowitz, *Methods of Numerical Integration*, Academic Press, New York, 1975.
33. G. A. Deschamps, and H. S. Cabayan, Antenna synthesis and solution of inverse problems by regularization methods. *IEEE Trans. Anten. Proc.* **20** (1972), 268–274.
34. C. L. Dolph, A current distribution for broadside arrays which optimizes the relationship between beam width and side–lobe level. *Proc. IRE* **34** (1946), 335–348.
35. D. G. Dudley, *The Mathematical Foundations for Electromagnetic Theory*, IEEE Press, New Jersey, 2001.
36. N. Dunford, and J. Schwartz, *Linear Operators, Part I*, John Wiley & Sons, New York, London, Sydney, 1957,

37. H. ENGL, Discrepancy principles for Tikhonov regularization of ill–posed problems leading to optimal convergence rates. *J. Optim. Theory Appl.* **52** (1987), 209–215.
38. R. L. FANTE, AND J. T. MAYHAN, Bounds on the electric field outside a radiating system. *IEEE Trans. Antennas Prop.* **16** (1968), 712–717.
39. R. L. FANTE, AND J. T. MAYHAN, Bounds on the electric field outside a radiating system–II. *IEEE Trans. Antennas Prop.* **18** (1970), 64–68.
40. S. FAST, *An Optimization Method for Solving a Radiation Direction Problem*, Ph.D. Thesis, Department of Mathematical Sciences, University of Delaware, Newark, DE, 1988.
41. D. GILBARG, AND N. S. TRUDINGER, *Elliptic Partial Differential Equations of Second Order*, Springer–Verlag, Berlin, Heidelberg, New York, 1977.
42. B. GIESEKE, Zum Dirichletschen Prinzip für selbstadjungierte elliptische Differentialoperatoren. *Math. Z.* **68** (1964), 54–62.
43. C. I. GOLDSTEIN, The finite element method with non–uniform mesh sizes applied to the exterior Helmholtz problem. *Numer. Math.* **38** (1981), 61–82.
44. C. I. GOLDSTEIN, The solution of exterior interface problems using a variational method with Lagrange multipliers. *J. Math. Anal. and Appl.* **97** (1983), 480–508.
45. J. HADAMARD, *Lectures on the Cauchy Problem in Linear Partial Differential Equations*, Yale University Press, New Haven, 1923.
46. R. E. HARRINGTON, AND J. R. MAUTZ, An impedance sheet approximation for thin dielectric shells. *IEEE Trans. Ant. Prop.* **23** (1975), 531–534.
47. M. R. HESTENES, *Optimization Theory. The Finite Dimensional Case.*, J. Wiley & Sons. New York, London, Sydney, Toronto, 1975.
48. H. HEUSER, *Funktionalanalysis*, Teubner–Verlag, Stuttgart, 1992.
49. H. HEUSER, *Lehrbuch der Analysis, Band 2*, Teubner–Verlag, Stuttgart, 2000.
50. H. HOCHSTADT, *The Functions of Mathematical Physics*, John Wiley & Sons, New York, London, Sydney, 1971.
51. R. B. HOLMES, *Geometric Functional Analysis and it Applications*, Springer–Verlag, Berlin, Heidelberg, New York, 1975.
52. S. R. HOLSTON, Optimization of multiple antenna array performance measures using a multicriteria approach. *IEEE Trans Antennas Prop.*, *submitted for publication.*
53. G. C. HSIAO, Mathematical foundations for the boundary-field equation methods in acoustic and electromagnetic scattering. In: *Analytical and Computational Methods in Scattering and Applied Mathematics*, F. Santosa and I. Stakgold eds., Chapman and Hall/CRC, Boca Raton, London, New York, Washington, D.C., 2000.
54. F. IHLENBURG, *Finite Element Analysis of Acoustic Scattering*, Springer Verlag, New York, 1998.
55. V. K. IVANOV, On linear problems which are not well–posed. *Soviet Math. Dokl.* **4** (1962), 981–983 (English translation).
56. J. JAHN, *Mathematical Vector Optimization in Partially Ordered Linear Spaces*, Peter Lang, Frankfurt, 1986.
57. K. JÖRGENS, *Linear Integral Operators*, Teubner–Verlag, Pittman Press, London, 1982.
58. C. JOHNSON, AND J. NEDELEC, On the coupling of the boundary integral and finite element methods. *Math. Comp.* **35** (1980), 1063–1079.

59. D. S. Jones, *Methods in Electromagnetic Wave Propagation, 2nd. ed.*, Clarendon Press, Oxford, 1994.
60. A. Jüschke, J. Jahn, and A. Kirsch, A bicriterial optimization problem of antenna design. *Comp. Optimiz. Appl.* **7** (1997), 261–276.
61. J. L. Kelley, *General Topology.* Springer Verlag, Berlin, Heidelberg, New York, 1991.
62. H. Kersten, Grenz– und Sprungrelationen für Potentiale mit quadratsummierbarer Dichte. *Resultate d. Math.* **3** (1980), 17–24.
63. H. Kersten, Die C–Vollständigkeit partikulärer Lösungssysteme der Schwingungsgleichung $\Delta U + k^2 U = 0$. *Res. d. Math.* **4** (1981), 155–170.
64. H. Kersten, Ein neuer Zugang zum ersten Randwert–Problem der Schwingungsgleichung in Gebieten mit nicht–glattem Rand. *Habilitation thesis*, Aachen, 1983.
65. A. Kirsch, The Robin problem for the Helmholtz equation as a singular perturbation problem. *Numer. Funct. Anal. and Optimiz.* **8** (1985), 1–20.
66. A. Kirsch, Remarks on some notions of weak solutions for the Helmholtz equation. *Appl. Anal.* **47** (1992), 7–24.
67. A. Kirsch, *An Introduction to the Mathematical Theory of Inverse Problems*, Springer–Verlag, Berlin, Heidelberg, New York, 1996.
68. A. Kirsch, Characterization of the scattering obstacle by the spectral data of the far field operator. *Inverse Problems* **14** (1998), 1489–1512.
69. A. Kirsch, and P. Monk, Convergence analysis of a coupled finite element and spectral method in acoustic scattering. *IMA J. Numer. Anal.* **10** (1990), 425–447.
70. A. Kirsch, and P. Monk, An analysis of the coupling of finite element and Nyström methods in acoustic scattering. *IMA J. Numer. Anal.* **14** (1994), 523–544.
71. A. Kirsch, W. Warth, and J. Werner, *Notwendige Optimalitätsbedingungen und ihre Anwendung* Lecture Notes In Economics And Mathematical Systems **v. 152**, Springer Verlag, Berlin, Heidelberg, New York, 1978.
72. A. Kirsch, and P. Wilde, The optimization of directivity and signal–to–noise ratio of an arbitrary antenna array. *Math. Meth. Appl. Sci.* **10** (1988), 153–164.
73. M. A. Krasnosel'skii, G. M. Vainikko, P .P. Zabreiko, Ya. B. Rutitskii, and V. Ya. Stetsenko, *Approximate Solution of Operator Equations* (English translation), Wolters–Noordhoff Publishing Company, Groningen, 1972.
74. R. Kress, *Linear Integral Equations*, Springer–Verlag, Berlin, Heidelberg, New York, (2nd edition), 1999.
75. R. Kress, A Nyström method for boundary integral equations in domains with corners. *Numer. Math.* **58** (1990), 145–161.
76. R. Kress, *Numerical Analysis*, Springer–Verlag, Berlin, Heidelberg, New York, 1998.
77. A. Kriegsmann, and C. S. Morawetz, Solving the Helmholtz equation for exterior problems with variable index of refraction:I. *SIAM J. Sci. Stat. Comput* **1** (1980), 371–385.
78. R. Kussmaul, Ein numerisches Verfahren zur Lösung des Neumannschen Aussenraumproblems für die Helmholtzsche Schwingungsgleichung. *Computing* **4** (1969), 246–273.
79. P. Kwok, and P. Brandon, Maximisation of signal/noise ratio in arrays with broadened zero., *Electron. Lett.*16 (1980), 60–62.

80. M. LASSAS, AND E. SOMERSALO, On the existence and convergence of the solution of the PML equations. Preprint, 1997.
81. E. LEPELLARS, AND T. S. ANGELL, A multicriteria optimization problem for a circular array of dipoles. *to appear.*
82. N. LIMIĆ, Galerkin–Petrov method for Helmholtz equation exterior problems. *Glasnik Matematicki* **16** (1981), 245–260.
83. N. LIMIĆ, The exterior Neumann problem for the Helmholtz equation. *Glasnik Matematicki* **16** (1981), 51–64.
84. I. V. LINDELL, *Methods for Electromagnetic Field Analysis*, IEEE Press, New Jersey, 1995.
85. Y. T. LO, S. W. LEE, AND Q. H. LEE, Optimization of directivity and signal–to–noise ratio of an arbitrary antenna array. *Proc. IEEE* **54** (1966), 1033–1045.
86. P. LORRAIN, D. R. CORSON, AND F. LORRAIN, *Electromagnetic Fields and Waves*, Freeman and Company, New York, 1988.
87. C. F. LOZANO, AND R. REEMTSEN, On a Stefan problem with an emerging free boundary. *Num. Heat Transfer* **4** (1981), 239–245.
88. D. G. LUENBERGER, *Optimization by Vector Space Methods*, John Wiley & Sons, New York, 1969.
89. D. MARGETIS, G. FIKORIS, J. M. MYERS, AND T. T. WU, Highly directive current distributions: General theory. *Physical Review E* **58** (1998), 2531–2547.
90. W. MAGNUS, AND F. OBERHETTINGER, *Formulas and Theorems for the Functions of Mathematical Physics*, J. Wermer tr., Chelsea Publishing Col, New York, 1954.
91. R. MACCAMY, AND S. MARIN, A finite element method for exterior interface problems. *Internat. J. Math & Math. Sci.* **3** (1980), 311–350.
92. S. MARIN, Computing scattering amplitudes for arbitrary cylinders under incident plane waves. *IEE Trans. on AP* **AP–30** (1982), 1045–1049.
93. E. MARTENSEN, Über eine Methode zum räumlichen Neumannschen Problem mit einer Anwendung auf torusartige Berandungen. *Acta Math.* **109** (1963), 75–135.
94. E. MARTENSEN, *Potentialtheorie*, B. G. Teubner, Stuttgart, 1968.
95. M. MASMOUDI, Numerical solutions for exterior problems. *Numer. Math.* **51** (1987), 87–101.
96. V. P. MIKHAILOV, On the boundary values of solutions of elliptic equation. *Applied Math. Optimization* **6** (1980), 193-199.
97. R. F. MILLAR, The Raleigh hypothesis and a related least squares solution to scattering problems for periodic surfaces and other scatterers. *Radio Science* **8** (1973), 785-796.
98. M. MINOUX, *Mathematical Programming: Theory and Algorithms*, John Wiley & Sons, 1986.
99. C. MIRANDA, *Partial Differential Equations of Elliptic Type*, 2nd. edition, Springer Verlag, Berlin, 1970.
100. P. MONK, *Finite Element Methods for Maxwell's Equations*, Oxford University Press, Oxford, 2003.
101. V. A. MOROZOV, Choice of parameter for the solution of functional equations by the regularization method. *Sov. Math. Dokl.* **8** (1967), 1000–1003.
102. V. A. MOROZOV, The error principle in the solution of operational equations by the regularization method. *USSR Comput. Math. Math. Phys.* **8** (1968), 63–87.
103. C. MÜLLER, Radiation patterns and radiation fields. *J. Rat. Mech. Anal.* **4** (1955), 235–246.

104. C. MÜLLER, Boundary values and diffraction problems. *Symposium Mathematica* **18** (1976), 354–367.

105. C. MÜLLER, *Foundations of the Mathematical Theory of Electromagnetic Waves*, Springer–Verlag, Berlin, Heidelberg, New York, 1969.

106. C. MÜLLER, AND H. KERSTEN, Zwei Klassen vollständiger Funktionensysteme zur Behandlung der Randwertaufgaben der Schwingungsgleichung $\Delta U + k^2 U = 0$. *Math. Meth. in the Appl. Sci.* **2** (1980), 48–67.

107. M. Z. NASHED, Generalized inverses, normal solvability, and iteration for singular operator equations. in *Nonlinear Functional Analysis and Applications* L. B. Rall, ed., 311–359, Academic Press, New York, 1971.

108. A. W. NAYLOR, AND G. R. SELL, *Linear Operator Theory in Engineering and Science*, Springer Applied Mathematical Sciences Series, Vol. 40, Springer Verlag, Berlin, Heidelberg, New York, 2000.

109. E. NICOLAU, AND D. SAHARIA, *Adaptive Arrays*, Elsevier, Amsterdam, Osford, New York, Tokyo, 1989.

110. A. G. RAMM, Optimal solution of the problem of linear antenna synthesis. *Sov. Phys. Dokl.* **13** (1968), 546–54.

111. R. REEMTSEN, AND S. GÖRNER, Numerical Methods for Semi–Infinite Programming: A Survey. In: *Semi–Infinite Programming*, R. Reemtsen and Rückmann eds., 1998, Kluwer Academic Publishers, 195–275.

112. R. REEMTSEN, AND A. KIRSCH, A method for the numerical solution of the one–dimensional inverse Stefan problem. *Numer. Math.* **45** (1984), 253–273.

113. F. RELLICH, Über das asymptotische Verhalten der Lösungen von $\Delta u + \lambda u = 0$ in unendlichen Gebieten. *Jber. Deutsch. Math. Verein.* **53** (1943), 57–65.

114. D. R. RHODES, The optimum line source for the best mean–square approximation to a given radiation pattern. *IEEE Trans. Antennas Prop* **11** (1963), 440–446.

115. D. R. RHODES, *Synthesis of Planar Antenna Sources*, Clarendon Press, Oxford, 1974.

116. G. SANTHOSH, AND M. THAMBAN NAIR, A class of discrepancy principles for the simplified regularization of ill–posed problems. *J. Austr. Math. Soc. Ser. B* **36**, (1995), 242–248.

117. A. H. SCHATZ, An observation concerning Ritz–Galerkin methods with indefinite bilinear forms. *Math. Comp.* **28** (1974), 959–962.

118. S. A. SCHELKUNOFF, A mathematical theory of arrays. *Bell Sys. Tech. Jour.* **22** (1943), 80–107.

119. R. SCOTT, Interpolated boundary conditions in the finite element method. *SIAM J. Numer. Anal.* **12** (1975), 404–427.

120. T. B. A. SENIOR, Impedance boundary conditions for imperfectly conducting surfaces. *Appl. Sci. Res. B*, **8** (1960), 418–436.

121. T. B. A. SENIOR, Backscattering from resistive strips. *IEEE Trans. Ant. Prop.* **27** (1979), 808–813.

122. T. B. A. SENIOR, AND M. NAOR, Low frequency scattering by a resistive strip. *IEEE Trans. Ant. Prop.* **32** (1984), 272–275.

123. R. A. SHORE, Sidelobe sector nulling with minimized weight perturbations, *RADC–TR–86–40*, ROme Air Development Center, Airforce Systems Command, (1985).

124. D. SLEPIAN, AND H. O. POLLAK, Prolate spheroidal wave functions, Fourier analysis and uncertainty–I. *Bell Syst. Tech. J.* **40** (1961), 43–64.

125. A. Sommerfeld, *Partial Differential Equations in Physics*, Academic Press, New York, 1957.

126. W. Stadler, A survey of multicriteria optimization or the vector maximization problem, I. *J. Optim. Theory Appl.* **29** (1979), 1–52.

127. W. Stadler, Multicriteria optimization in mechanics (a survey). *Appl. Mech. Rev.* **37** (1984), 277–286.

128. H. Steyskal, Synthesis of antenna patterns with prescribed nulls. *IEEE Trans. Antennas Propag.* **AP–30** (1982), 273–279.

129. G. Still, On density and approximation properties of special solutions of the Helmholtz equation. *ZAMM* **72** (1992), 277–290.

130. J. A. Stratton, *Electromagnetic Theory*, McGraw Hill, New York, London, 1941.

131. J. W. Strutt (Lord Rayleigh), On the dynamical theory of gratings. *Proc. Roy.Soc.* **79** (1907), 339-416.

132. W. L. Stutzman, and G. A. Thiele, *Antenna Theory and Design, 2nd. ed.*, John Wiley & Sons, Inc., New York, 1998.

133. T. T. Taylor, Design of line–source antennas for narrow beamwidth and low side lobes. *IRE Trans. Antennas Prop.* **3** (1955), 16–28.

134. A. N. Tikhonov, On the stability of inverse problems. *Dokl. Akad. Nauk SSSR* **39** (1943), 195–198 (in Russian).

135. A. N. Tikhonov, Regularization of incorrectly posed problems. *Sov. Math. Doklady* **4** (1963), 1624–1627.

136. F. Treves, *Topological Vector Spaces, Distributions and Kernels*, Academic Press, New York, 1967.

137. F. G. Tricomi, *Integral Equations*, Wiley Interscience, New York, London, Sydney, 1967.

138. M. M. Vainberg, *Variational Methods for the Study of Nonlinear Operators*, Holden–Day, San Francisco, 1964.

139. G. N. Watson, *A Treatise on the Theory of Bessel Functions*, Cambridge University Press, Cambridge, 1966.

140. D. S. Weile, and E. Michielssen, Integer coded Pareto genetic algorithm design of antenna arrys. *Electronics Letters***32** (1996), 1744–1755.

141. D. S. Weile, E. Michielssen, and D. E. Goldberg, Genetic algorithm design of Pareto optimal broad band microwave absorbers. *IEEE Trans. Electromag. Compat.* **45** (1996), 518–524.

142. L. Wheeden and A. Zygmund, *Measure and Integral; An Introduction to Real Analysis*, Marcel Decker Inc., New York, Basel 1977.

143. E. T. Whittaker, *A History of the Theories of Aether and Electricity from the Age of Descartes to the Close of the Nineteenth Century*, T. Nelson, New York, 1951.

144. J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Oxford University Press, Oxford, 1965.

145. K. Yosida, *Functional Analysis*, 6th. ed., Springer Verlag, Berlin Heidelberg New York, 1980.

Index