Fundamentals of Inverse Problems[1]

Eric L. Miller
Department of Electrical and Computer Engineering &
The Center for Subsurface Sensing and Imaging Systems
Northeastern University, Boston MA 02115
Phone: (617) 373-8386
FAX: (617) 373-8627
email: elmiller@ece.neu.edu
and
W. Clem Karl
Department of Electrical and Computer Engineering &
The Center for Subsurface Sensing and Imaging Systems
Boston University, Boston MA 02215
Phone: (617) 353-2811
FAX: (617) 353-1282
email: wckarl@bu.edu

September 7, 2003

To our families for their endless patience, smiles, and love.

Contents

1 Introduction and Caveats **4**

2 Mathematical Preliminaries **8**
2.1 Linear Vector Spaces 8
2.2 Functionals and Operators 13
2.2.1 Functionals and Dual Spaces 14
2.2.2 Operators, Adjoints, and the Four Fundamental Subspaces 16
2.2.3 Eigenanalyis and the Singular Value Decomposition 20
2.3 Exercises 28

3 Forward Models & Inverse Problems **32**
3.1 Deconvolution 32
3.2 X-ray Tomography 34
3.3 Inverse Source and Inverse Scattering Problems 35
3.3.1 The Helmholtz Model 35
3.3.2 Green's Functions 37
3.3.3 The Inverse Problems 40
3.4 Discretization Methods 45
3.5 Exercises 56

4 Analytic Methods for Linear Inverse Problems **68**
4.1 Inverting the Radon Transform 70
4.2 Diffraction Tomography: Inverting the Born Approximation 72
4.3 Exercises 82

5 Numerical Methods for Linear Inverse Problems **84**
5.1 Ill-posedness 85
5.1.1 Existance 86
5.1.2 Uniqueness 87
5.1.3 Stability 89
5.2 The Pseudo-Inverse 98
5.2.1 Full-Rank Overdetermined Inverse Problems 98
5.2.2 Full-Rank Underdetermined Inverse Problems 100

5.2.3 Reduced Rank Problems . . . 101
5.3 Regularization I . . . 111
5.3.1 The Truncated Singular Value Decomposition . . . 112
5.3.2 Spectral Filtering . . . 113
5.3.3 Variational Regularization Methods . . . 115
5.3.4 Tikhonov-type Methods . . . 117
5.3.5 Parameter Selection . . . 119
5.3.6 Examples . . . 124
5.4 Semi-discrete Linear Inverse Problems . . . 124
5.5 Exercises . . . 125

6 Numerical Methods for Nonlinear Inverse Problems 131
6.1 A Review of Optimization Theory and Algorithms . . . 132
6.1.1 General Unconstrained Problems . . . 133
6.1.2 Nonlinear Least Squares Problems . . . 136
6.2 Regularization II: Edge Preservation . . . 137
6.3 Nonlinear physical models . . . 142
6.3.1 Adjoint Field Calculations . . . 142
6.3.2 Integral Equation Method . . . 145
6.4 Geometric Inverse Methods . . . 148
6.5 Exercises . . . 148

A A Brief Review of Probability 151
A.1 Basic Concepts . . . 151
A.2 Random Variable . . . 152
A.3 Jointly Distributed Random Variables . . . 155

Chapter 1

Introduction and Caveats

We consider inverse problem in a very broad sense: the desire to extract information regarding an unknown quantity based on data related via some type of mathematical model to the unknown. In most all cases of interest here, this model (known as a *forward model*) is derived from physical principles relating the data to the desired quantity. Whether explicitly accounted for or not as part of the processing, the data are typically corrupted by some form of noise. The desired quantity may be a function of space, time, and/or frequency. The sought-after information is also quite varied depending on the underlying application. One may require a detailed reconstruction of the internal structure of a medium; i.e, an image in two dimensions or a volumetric rendering in three. Such is the case in the vast majority of image restoration problems as well many problems arising in medical imaging. Alternatively, the goal may be the direct characterization of anomalous areas in a larger field of regard. For example, in non-destructive test and evaluation, one may not be concerned with the nominal structure of a sample, but only with knowing whether there exist flaws and if so where they are located and their morphologies. In these cases and in contrast to imaging problems where millions of pixel/voxel values may be desired, the required information here may be nothing more than a small collection of parameters defining the number, location, and shape characteristics of the anomalies. In more complicated problems, these images and/or objects are temporally dynamic thereby adding an additional dimension to the problem. Finally we note there is growing interest for solving inverse problems where multiple quantities are simultaneously to be determined. The use of ultrasound to recover maps of sound speed *and* acoustic attenuation is one example. Optical tomographic techniques can be of use in extracting information regarding the spatio-temporal dynamics of a number of chromophores (oxygenated hemoglobin, de-oxygenated hemoglobin, lipids, water, ...) in the human body.

Given the breadth associated just with the simple definition of an inverse problem, it should come as no surprise that there are a plethora of methods that have been developed in a wide range of intellectual communities for solving these problems. In many important cases, mathematically exact formula can be used as the basis for "inverting" the forward model to obtain the desired information. Perhaps the best known example here is the use of the filtered backprojection (FBP) method [47] as applied in computer aided tomography (CAT) and magnetic resonance imaging (MRI). Closely related is the filtered bakpropagation approach [23] appropriate for problems in fields such as geophysics where diffractive effects are of import. Finally, we

mention the more mathematically sophisticated and highly elegant techniques largely under development in the mathematics and mathematical physics communities including linear sampling [19], $\bar{\partial}$ (D-bar) [82],micro-local tomography [16], and dual space methods [20]. These inverse methods also provide analytic (or nearly analytic) formulae and algorithms for directly obtaining images or object information from data where the underlying physical model is more complex than that associated with filtered bakprojection or backpropagation.

At the other end of the inverse problems spectrum are variational methods which rely on numerical optimization for providing the required information. The recovered collection of voxel values or geometric parameters is defined as that collection of quantities which minimize a cost function [7, 22]. The cost function itself is usually comprised of (a) a term requiring that the extracted information in some sense be consistent with the measured data and (b) a number of additional terms used to enforce prior information one may possess concerning properties of the solution. For example, in the case of image formation problems, such terms usually enforce a degree of smoothness in the reconstructed imagery [13]. The final cost function then represents a balance between these two competing sources of information: data and prior knowledge. Actually determining the minimizer of the cost function is the accomplished through the use of numerical methods drawn from the optimization community. Thus, in contrast to the analytical techniques described in the previous paragraph which find a home in the mathematical and math physics communities, those of interest here are more commonly the purview of engineers, applied mathematicians and applied physicists.

It almost goes without saying that the study and practice of inverse problems is not nearly as neatly drawn as the last two paragraphs might indicate. The rapid rise in multi-disciplinary research in the last 15 years or so has seen a concomitant merging of the two classes of methods defined above: the analytical and the algorithmic. While there certainly exists cutting edge work in each separately, it would not be a mis-statement to say that the fusion of these techniques is itself an area of active work.

Perhaps the most telling tale that can be gleaned from the coarse taxonomy presented here is that inverse problems represents an exceptionally broad area of study which can draw from a wide range of disciplines. Thus there are many possible directions from which a text on such a topic can approach the subject matter. The tact taken here is aimed at satisfying the interests of a mathematically savvy researcher who ultimately wants an algorithm that can be implemented on a computer. The methods and models addressed here are placed within a vector space/operator theoretic framework; however ours is by no means a rigorous functional theoretic approach to the study of inverse problems. Wonderful texts in this category exist such as [28, 49, 69] and it is neither the goal nor within the author's expertise to add to this literature. I view the mathematics as providing a useful and elegant language by which common issues arising across a range of application and problem structures can be conveniently discussed. To this end, the first chapter of this manuscript is devoted to a brief overview of the relevant linear vector space mathematics required for the remainder.

In order to present inverse problems in any comprehensive manner, one must first understand the underlying forward models. Chapter 3 provides a review of the physical models and related inverse problems to be encountered in the remainder of the text. Four classes of problems shall be considered: deconvolution, X-ray tomography, and frequency-domain inverse source and inverse

scattering scattering associated with the scalar Helmholtz equation. On the assumption that anyone reading these notes is familiar with basic linear systems theory, deconvolution provides a natural forward model whose associated inverse problems are very easily understood using basic tools from Fourier analysis. X-ray tomography (or the Radon transform) is one step up in complexity from convolution. The utility of Fourier methods in the study of this problem and its ubiquity especially in medical imaging applications makes it well suited for the analytic-algorithmic balance that is the focus of this text. Inverse source and scattering problems encountered with the Helmholtz equation are a step up in complexity and generality again from X-ray tomography. Problems for which the scalar Helmholtz equation is a valid and useful model abound: medical imaging, nondestructive evaluation, geophysical prospecting, environmental remediation, remote sensing and surveillance to name but a few. Just as broad are the classes of instruments and sensors encountered in these applications whose physics are described by this model: acoustics, scalar electromagnetic (including as examples DC resistivity tools, low frequency induction sensors, microwave and RF ground penetrating radar, up to terahertz imaging tools), photo-thermal, and diffuse optical. The physics of these problems are significantly more complicated and rich than convolution or the Radon transform. The inverse source problem end up being linear and quite closely related to deconvolution. The inverse scattering problem is nonlinear; however linearization produces structure very much analogous to the X-ray case.

Given this background, we next turn our attention to linear inverse problems. By linear inverse problems we really mean problems whose variational forms can put into some type of linear least squares structure. Such problems have two components. First, the forward model must be linear. Convolution, the Radon transform, and linearized inverse scattering fall into this category. In fact, these are precisely the problems for which elegant and useful analytic inversion formulae exist. Second, when dealing with regularization in the context of variational methods, the regularizer must be quadratic in the desired quantities. These apparent restrictions however still provides for significant interesting structure and form the basis for a great majority of the inverse methods currently in practical use. Indeed, this is not a coincidence. Inverse problems which can be connected to linear least squares formulations are far easier to study than their non-linear cousins. Analytic inversion formulae are more easily obtained for broad classes of problems encountered in practice. Closed form solutions exist for the variational formulations of many of these problems. Finally, most all of the tools used to solve these problems will be encountered in the study of nonlinear problems.

After thoroughly looking at linear inverse problems, we come to their non-linear counterparts in Chapter 6. Nonlinearity in this case arises from one of two sources: either the model linking the unknowns to the data is not linear in the parameters to be recovered or the regularization is not quadratic in these quantities. All geometric inverse methods where we are concerned with things like size, shape, orientation, and number of anomalies fall into this category. In this text we concentrate on variational approaches to solving these problems. Generally-applicable decent type of optimization methods will be discussed. At the heart of these techniques is the need to compute the gradient of the cost function with respect to the unknowns. Methods for accomplishing this task will be described again using convolution, X-ray tomography, and inverse scattering as the driving examples. Finally, we shall briefly discuss current work in the development of closed form methods for solving an important subset of these problems.

Having stated all of this, it is time now for the caveats. It is important to know what is not

covered in this manuscript. The short answer is, "Many things." Connections between variational approaches to inverse problems and statistical inference and estimation are not described here. Problems in which the underlying phenomena have a vectorial nature such as electromagnetics using the full Maxwell's equation or elasticity are also reserved for more advanced texts on inverse methods. While there are many important areas where such physics represent the sensors, the level of detail and additional mathematical machinery required to address these problems put them, in my opinion, outside of the scope of anything that could be construed as an introductory treatment of inverse methods. The same holds true for problems whose underlying physics are described by nonlinear partial differential equations as arise in many facets of acoustics. Finally, as noted previously, I have decided to leave the detailed and mathematically rigorous functional-theoretic and harmonic-analytic approaches to the study of inverse problems to other, far more qualified authors.

The other caveat is associated with the nature of this document. In its current form, this manuscript represents a transcribed version of a set of lecture notes used by the author in the teaching of a graduate level class at Northeastern University. As time and resources permit, it is the hope and desire of the author to add all of the required connective tissue (text, pictures, problems, references, an index, ...) to these notes which will one day allow them to be called a book. In that spirit then, all comments, correction, and suggestions should be emailed to the authors at `elmiller@ece.neu.edu` and `wckarl@bu.edu`.

Chapter 2

Mathematical Preliminaries

To establish notation and ensure a consistent understanding of the mathematical bases by the reader, we begin with a review of linear vector spaces and basic notions of linear operators. Much of the material in this text exists in the confines of finite dimensional spaces (matrices and vectors) thus perhaps leading one to wonder as to the necessity of the extra mathematical baggage. This is not intended to be, nor is it likely to be mistaken for, a tome of functional theoretic methods in the analysis and solution of inverse problems. Still the restriction of the topic to purely finite dimensional spaces is too limiting both because the underlying problems arise in these more mathematically abstract contexts and because there exists a range of solution methods that can be pursued in these domains whose forms and structures are so similar to the finite dimensional case and so grapsable by readers with an engineering background that it would be a shame to not present them.

Thus, here we pursue an engineers solution to the problem of balancing mathematical rigor against practical utility. Much as basic continuous time signal processing can make use of abstractions like a Dirac delta function without the need to take first a class in distribution theory, here we make use of ideas such as Banach and Hilbert spaces, operator adjoints, eigenfunctions and singular functions without much concern for the subtle and complex issues associated with these topics that is central to their rigorous study. For more extensive discussions of these and related topics, the interested reader is referred to [52, 55] which provided much of the motivation for the material presented here.

2.1 Linear Vector Spaces

Definition 2.1 *A* linear vector space *(or just vector space), X, is a set of elements (called vectors or functions) closed under two operations, addition and scalar multiplication:*

1. *Addition: For $x \in X$ and $y \in X$, $x + y \in X$.*
2. *Scalar multiplication: For α a complex number (denoted as $\alpha \in \mathbb{C}$) and $x \in X$, $\alpha x \in X$.*

While there mare many flavors of linear vector spaces, here we deal with the nicest possible kind where the following properties are all assumed to hold for $x, y, z \in X$ and $\alpha, \beta \in \mathbb{C}$:

1. Commutivity: $x + y = y + x$
2. Associativity: $(x + y) + z = x + (y + z)$
3. Existance of a zero vector and a unit vector: There is an element $0 \in X$ such that $x + 0 = x$.
4. Distributivity I: $(\alpha + \beta)x = \alpha x + \beta x$
5. $(\alpha\beta)x = \alpha(\beta x)$
6. For the scalars 0 and 1, $0x = 0$ and $1x = x$.

Examples of vector spaces where these properties hold include the following[1]:

Example 2.1 The real numbers, $\mathbb{R}$

Example 2.2 Standard n dimensional Euclidean spaces, $\mathbb{R}^n$ whose elements are taken to be column vectors of the form

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

where the x_i are real valued scalars.

Example 2.3 Semi-infinite length sequences $x = (x_1, x_2, x_3, \ldots, x_j, x_{j+1}, \ldots)$. Two classes of such subspaces are those sequences whose first p elements are identically zero and those sequences whose values converge to zero as $j \to \infty$.

Example 2.4 All real-valued, continuous functions, $f(x)$ on the interval $x \in [0, 1]$

Definition 2.2 *The set M is a* subspace *of a linear vector space if for all $x, y \in M$ and $\alpha, \beta \in \mathbb{C}$, it is the case that $\alpha x + \beta y \in M$ as well.*

Definition 2.3 *If S and T are subsets of X, the sum of S and T is the collection of all vectors of the form $s + t$ where $s \in S$ and $t \in T$.*

From these last two definitions one can prove that if M and N are subspaces of a linear vector space X, then $M + N$ is a subspace of X as well.

Definition 2.4 *A* linear combination *of the vectors $x_1, x_2 \ldots, x_n$ is a sum of the form $\alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n$ for complex scalars α_i.*

Definition 2.5 *A vector x is said to be* linearly dependent *on a set of vectors S if x can be written as a linear combination of vectors from S. Otherwise, x is said to be* linearly independent *of the elements of S. A set of vectors is said to be linearly independent if each is linearly independent from the others.*

[1] EXERCISE: Show these are vector spaces

From these last two definitions, one can prove that $x_1, x_2, \ldots, x_n$ are linearly independent if and only if

$$\sum_{i=1}^{n} \alpha_i x_i = 0$$

hold only for all $\alpha_i = 0$. Examples of linear independent and dependent vectors in $\mathbb{R}^2$ are shown in Fig. 2.1. For the case of linear dependence, choosing the three α_i's identically equal to 1 will yield the zero vector.

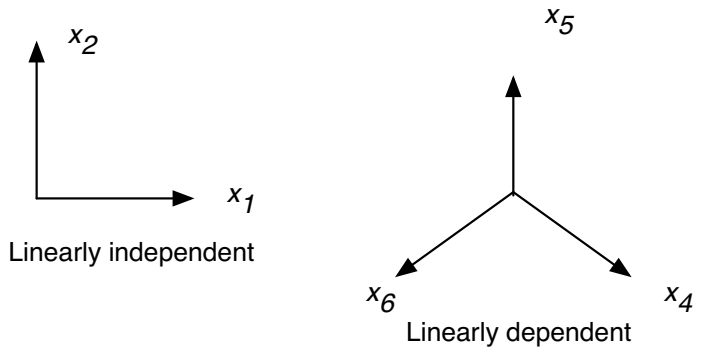


Figure 2.1: Linearly independent and dependent vectors in $\mathbb{R}^2$

Definition 2.6 *A finite set of linearly independent vectors $x_1, x_2, \ldots x_n$ is said to be a* basis *for the vector space X is any $s \in X$ can be written as a linear combination of the x_i.* Finite *dimensional spaces are those whose bases are comprised of a finite number of vectors. Otherwise, the space is termed* infinite dimensional.

Definition 2.7 *A* normed linear vector space *is a vector space, X, equipped with a real valued function called a* norm *satisfying the following three properties for $x, y \in X$ and $\alpha \in \mathbb{C}$*

1. *Non-negativity: $\|x\| \geq 0$ with equality if and only if $x = 0$.*
2. *Triangle inequality: $\|x + y\| \leq \|x\| + \|y\|$*
3. *$\|\alpha x\| = |\alpha| \|x\|$*

Some examples of normed linear vector spaces include[2]:

Example 2.5 Continuous functions $f(t)$ on the interval $t \in a, b$ with the norm

$$\|f\| = \max_{a \leq t \leq b} |f(t)|$$

[2]EXERCISE

Example 2.6 Sequences of the form $x = \{\zeta_1, \zeta_2, \zeta_3, \ldots, \zeta_n, 0, 0, 0 \ldots\}$ with the norm

$$\|x\| = \sum_{i=1}^{n} |\zeta_i|$$

Example 2.7 Continuous functions $f(t)$ on the interval $t \in a, b$ with the norm

$$\|f\| = \int_a^b |f(t)| dt$$

Such spaces are denoted $C[a, b]$.

Definition 2.8 *A* Cauchy sequence *is a set of vectors $x_1, x_2, \ldots$ such that $\|x_n - x_m\| \to 0$ as $m, n \to \infty$. Formally, this means that for all $\epsilon > 0$ there is an N such that $\|x_n - x_m\| < \epsilon$ for $m, n > N$.*

Definition 2.9 *A vector space is* complete *if every Cauchy sequence converges to a point in that space.*

Definition 2.10 *A complete, normed, linear vector space is called a* Banach space.

Examples of Banach spaces include:

Example 2.8 $C[0, 1]$

Example 2.9 The l_p spaces consist of those sequences $x = \{\zeta_1, \zeta_2, \zeta_3, \ldots\}$ such that

$$\|x\| \equiv \left(\sum_{i=1}^{\infty} |\zeta_i|^p \right)^{1/p} < \infty$$

with $\|x\|_\infty = \sup_i |\zeta_i|$

Example 2.10 The $L_p[a, b]$ spaces [3] are sets of functions $f(t)$ for which

$$\|f\| \equiv \left(\int_a^b |f(t)|^p dt \right)^{1/p} < \infty$$

with $\|f\|_\infty = \operatorname{ess\,sup}_t |x(t)|$ where the essential supremum is defined in XXX.

Definition 2.11 *A* pre-Hilbert *space is a linear vector space equipped with a complex valued function called an* inner product *defined on $X \times X$ and for $x, y \in X$ denoted $(x|y)$ which satisfies the following properties for $x, y, z \in X$ and $\alpha \in \mathbb{C}$*

1. *$(x|y) = \overline{(y|x)}$ with the bar indicating complex conjugation.*

[3] L for the French mathematician Lesbegue.

2. $(x + y|z) = (x|y) + (x|z)$
3. $(\alpha x|y) = \alpha(x|y)$
4. $(x|x) \geq 0$ *with equality if and only if* $x = 0$.

Definition 2.12 *Two vectors, x and y, in a pre-Hilbert space are said to be* orthogonal *if $(x|y) = 0$ in which case we write $x \perp y$.*

With these properties one can show

1. $\sqrt{(x|x)}$ is a norm
2. $|(x|y)| \leq \|x\| \, \|y\|$ with equality if and only if $y = 0$ or y is a scalar multiple of x.
3. $\|x + y\|^2 + \|x - y\|^2 = 2\left(\|x\|^2 + \|y\|^2\right)$
4. If $x \perp y$ then $\|x + y\|^2 = \|x\|^2 + \|y\|^2$.

Definition 2.13 *A complete pre-Hilbert space is called a* Hilbert space

Definition 2.14 *Let S be a subset of a pre-Hilbert space, H. The set of all vectors in H that are orthogonal to S is called the* orthogonal complement *of S in H and is denoted $S^{\perp}$.*

That is

$$S^{\perp} = \{x \in X | (x|s) = 0 \text{ for all } s \in S\}$$

With this definition of the orthogonal complement we can show that for S and T subsets of a pre-Hilbert space H the following properties hold:

1. $S^{\perp}$ is a closed subspace of H. Roughly this means that the limit points of all convergent series of elements in $S^{\perp}$ are themselves contained in $S^{\perp}$.
2. $S \subset \left(S^{\perp}\right)^{\perp}$
3. $S \subset T \rightarrow T^{\perp} \subset S^{\perp}$
4. $S^{\perp\perp\perp} = S^{\perp}$

Definition 2.15 *A vector space X is the* direct sum *of two other spaces X_1 and X_2 if all vectors $x \in X$ have a unique representation*

$$x = x_1 + x_2$$

with $x_1 \in X_1$ and $x_2 \in X_2$. In this case we write $X = X_1 \oplus X_2$.

Given this notion of a direct sum it is possible to show that if X_1 is a closed subspace of a Hilbert space H then two convenient properties hold:

1. $H = X_1 \oplus X_1^{\perp}$
2. $X_1^{\perp\perp} = X_1$

Definition 2.16 *For $x \in X$, we call any $s \in S$ for which $x - s \in S^{\perp}$ the* orthogonal projection *of x into S.*

A graphical illustration of orthogonal projection is provided in Fig. 2.2. Here the vector x lives in the whole space, $X = \mathbb{R}^3$. The closed subspace of interest is S which in this case is spanned by the vectors x_1 and x_2. The orthogonal projection of x into S is the vector $x - s$ where s is part of the orthogonal complement of S in X. In this simple picture S^{perp} is spanned by x_3.

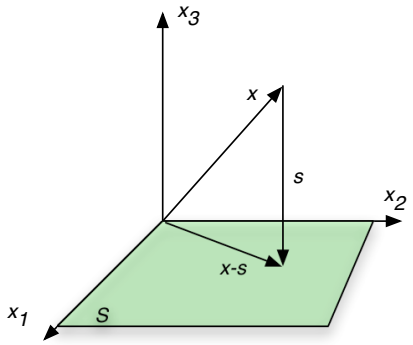


Figure 2.2: The orthogonal projection of a vector, x in $\mathbb{R}^3$ into the $x_1 - x_2$ plane is the vector $x - s$.

Armed with this collection of ideas all related to the notion of orthogonality, we come to a central result in the theory of linear vector spaces which in turn shall be used extensively in the study specifically of linear inverse problems.

Theorem 2.1 *(***The Projection Theorem** *[55, p. 51]) Let H be a Hilbert space and S a closed subspace of H. For any $x \in H$ there is a unique $\hat{s} \in S$ such that $\|x - \hat{S}\| \leq \|x - s\|$ for any $s \in S$. That is $\hat{s}$ is a solution to the following optimization problem*

$$\hat{s} = \arg\min_{s \in S} \|x - s\|.$$

Moreover, $\hat{s}$ is the minimizer if and only if $(x - \hat{s}) \perp S$.

2.2 Functionals and Operators

In this section we deal with mathematical constructs that represent generalizations of real valued functions. From an input-output perspective, a function may be thought of as a machine which takes as input a real number and produces as output again a real number. This idea is somewhat naturally extended in two ways. First, a functional takes as input a vector in some vector space, but, like a function, produces a simple real number. Second, an operator takes an element of a vector space and produces an element of a generally different vector space.

Both linear functionals and transformations have specialized forms commonly encountered in a technical undergraduate curriculum. If we take as our vector space $\mathbb{R}^n$ and elements of this space as

column vectors then one can obtain a real number from a vector through the use of a dot product. In this case "multiplication by a row vector" is the linear functional. Alternatively, production of a new vector $y \in \mathbb{R}^m$ (with m not necessarily equal to n) can be accomplished by left multiplication of $x \in \mathbb{R}^n$ with a matrix A having m rows and n columns. In this case, a transformation is represented by the matrix A and the action of taking x to y is achieved by matrix multiplication. Clearly for finite dimensional spaces, linear functionals are special cases of transformations where $m = 1$.

To obtain a slightly more general, but still hopefully accessible, example of a transformation, we can turn to the subject of linear time invariant systems in continuous time. It is well known that the output of such systems is given as the convolution of the input with the system's so-called impulse response. If we restrict the input signals to be finite in energy and assume the system is say causal and stable, the output signals will also be well behaved and we may view the system as a mechanism that takes as input one class of signals (functions, vectors, ...) and produces as output more signals (functions, vectors, ...). The transformation then is achieved via the convolution integral.

In the following sections (and really throughout most of this book), it should be helpful to fall back on the intuition gleaned from elementary linear algebra, signals and systems and related classes to provide examples of ideas that at first glance may appear abstract.

2.2.1 Functionals and Dual Spaces

Definition 2.17 *A* functional *is a mapping of $x \in X$ to a real number. Here x is a normed linear vector space.*

Symbolically, the action of a functional f on a vector x is denote by $f(x)$. Of particular interest to us for much of this book are *linear functions* defined as the following

Definition 2.18 *A* linear functional *is a functional satisfying $f(\alpha x + \beta y) = \alpha\, f(x) + \beta\, f(y)$ for $x, y \in X$ and $\alpha, \beta \in \mathbb{R}$.*

It turns out that linear functionals on a normed linear vector space X are themselves a vector space if we agree on the following two conventions:

1. $(f_1 + f_2)(x) = f_1(x) + f_2(x)$ for any two linear functionals f_1 and f_2 and $x \in X$.
2. $(\alpha f)(x) = \alpha f(x)$ for $\alpha \in R$, $x \in X$ and f a linear functional on X.

Example 2.11 Let X be the Euclidean space $\mathbb{R}^n$ so that any $x \in X$ can be written

$$x = \begin{bmatrix} \xi_1 \\ \xi_2 \\ \vdots \\ \xi_n \end{bmatrix}.$$

A linear functional on X takes the form

$$f(x) = \sum_{i=1}^{n} \eta_i \xi_i$$

and, depending on the situation, we write this operation as y^*x. That is, a linear functional on $\mathbb{R}^n$ is represented as a row vector $y = [\eta_1\ \eta_2\ \ldots\ \eta_n]$ and we can interpret the symbols y^*x as the common dot product between y and x.

Example 2.12 Consider now $L_2[0,1]$, the Hilbert space of square integrable functions on the interval $[0,1]$. A linear functional in this case is the continuous analog to a dot product; that is an integral against a weighting function. More formally, a linear functional on $L_2[0,1]$ is $f(x) = \int_0^1 y(t)x(t)dt$ for $y \in L_2[0,1]$.

Continuing with the case of $\mathbb{R}^n$, we know that linear functionals here take the form of row vectors again in $\mathbb{R}^n$. In other words, in simple Euclidean spaces, one may build a vector space out of the collection of linear functionals. This new space, called a *dual* is merely another copy of $\mathbb{R}^n$. It turns out that such a dual space can be constructed in more general circumstances than Euclidean spaces, but in order to make sure that they are well defined, the linear functionals have to be well behaved in the sense that the number they give as output cannot be infinity for any legal input $x \in X$. This lead then to the following:

Definition 2.19 *A* bounded linear functional *is a linear functional for which $|f(x)| < M\|x\|$ for some $M < \infty$ and any $x \in X$.*

The amount of "magnification" provided by a bounded linear functional is captured in its norm:

Definition 2.20 *The norm of a bounded linear functional, $\|f\|$, is defined in the following three equivalent ways:*

1. $\inf\left\{M \middle| |f(x)| < M\|x\| \text{ for all } x \in X\right\}$
2. $\sup_{x\neq 0} \frac{|f(x)|}{\|x\|}$
3. $\sup_{\|x\|=1} |f(x)|$

Now we can build a general version of a dual space.

Definition 2.21 *Let X be a normed linear vector space. The* normed dual *of X is denoted X^* and is defined to be the space of all bounded linear functionals on X.*

Given an element of the dual space, $x^* \in X^*$, the action of x^* on some $x \in X$ is represented as $x^*(x)$ or using angle brackets as $\langle x, x^*\rangle$. Some more examples should help further illustrate these ideas

Example 2.13 As we already know, the normed dual of $\mathbb{R}^n$ is again $\mathbb{R}^n$.

Example 2.14 Recall the definition of the sequence spaces l_p on page 11. The dual of l_p is l_q where $q = \frac{p}{1-p}$ i.e., $\frac{1}{p} + \frac{1}{q} = 1$. All bounded linear functionals on l_p take the form $y^*(x) = \sum_{i=1}^{\infty} \eta_i x_i$ with $x \in l_p$ and $y \in l_q$. Finally, the norm of $y \in l_q$ is

$$\|y\| = \left(\sum_{i=1}^{\infty} |\eta_i|^q\right)^{1/q}.$$

Example 2.15 Much as in the case of countable sequences, for the Lesbegue spaces, the dual of the space $L_p[0,1]$ is $L_q[0,1]$ where again $\frac{1}{p}+\frac{1}{q}=1$. Letting sums go to integrals gives:

$$\begin{aligned} y^*(x) &= \int_0^1 x(t)y(t)dt \\ \|y\| &= \left(\int_0^1 |f(t)|^q dt\right)^{1/q} \end{aligned}$$

Example 2.16 If H is any Hilbert space and f a bounded linear functional on H then it turns out there is a unique $y \in H$ such that for any $x \in H$, $f(x) = (x|y)$ and $\|f\| = \|y\|$

2.2.2 Operators, Adjoints, and the Four Fundamental Subspaces

While a functional is a mapping of an element of a vector space into a real number, a transformation (or operator) maps vectors in one space into vectors in another. That is, for some x an element of a normed linear vector space, $T(x)$ is an element y of another space Y. We write this in one of a number of ways including $T : X \to Y$ to indicate that T takes X into Y or $y = T(x)$. It is not necessarily the case that T is defined for all $x \in X$ nor is it necessarily the case that there exists and $x \in X$ such that $y = T(x)$ for any $y \in Y$. This leads us to the following definitions:

Definition 2.22 *The* domain *of T is that subset of X over which T is defined.*

Definition 2.23 *The* range *of T is the set of $y \in Y$ such that there exists an $x \in X$ for which $T(x) = y$. Occasionally we use $T(D)$ to indicate the range of the operator T where $D \subset X$ is the domain.*

Closely associated with the notions of domain and range are surjectivity, injectivity, and bijectivity:

Definition 2.24 *An operator $T : X \to Y$ is* surjective *(or* onto*) if for every $y \in Y$ there is an $x \in X$ such that $T(x) = y$.*

Definition 2.25 *An operator $T : X \to Y$ is* injective *(or* one-to-one*) if $T(x_1) = T(x_2)$ implies $x_1 = x_2$.*

Definition 2.26 *An operator which is both surjective and injective is said to be* bijective.

From a less formal perspective, a surjective transformation basically fills out all of Y as x varies over the domain, $D \subset X$. It may well be the case that there is more than one x which maps into any given y. In the event that each y in the range of T comes from only one x, then we have an injective operator. If an operator is both injective and surjective it means that every $y \in Y$ is uniquely associated with one and only one $x \in D$.

In the case that the operator satisfies linearity (i.e., $T(\alpha_1 x_1 + \alpha_2 x_2) = \alpha_1 T(x_1) + \alpha_2 T(x_2)$ for $x_i \in D$ and $\alpha_i \in \mathbb{C}$), then it turns out the range is more than just a set. It is in fact a subspace of Y and is denotes as $\mathcal{R}(T)$. Additionally for linear operators, there is a second subspace of intense interest

Definition 2.27 *The* nullspace *of a linear operator T, $\mathcal{N}(T)$, is defined as the set of all $x \in D$ for which $T(x) = 0$.*

As we discuss below, linear operators behave in most respects very much like matrices. Hence for this class of operators we adopt a more linear algebraic notation and replace $T(x)$ by Tx, reminiscent of the operation of matrix-vector multiplication.

In the same way that we define bounded linear functionals, bounded linear operators are those for which $\|Tx\| < M\|x\|$ for some $M < \infty$ and the induced norm of such operators is

$$\|T\| = \sup_{\|x\|=1} \|Tx\|.$$

The following are some examples of linear operators:

Example 2.17 Let $X = \mathbb{R}^n$ and $Y = \mathbb{R}^m$. Then a linear operator is nothing more than an $m \times n$ rectangular matrix, A, which maps n dimensional vectors into m dimensional vectors. In cooordinates we have:

$$y_i = \sum_{j=1}^{n} A_{i,j} x_j \quad i = 1, 2, \ldots m \tag{2.1}$$

with y_i the i-th element of the vector y, x_j the jth element of x and $A_{i,j}$ the element of A on row i and column j. If $m > n$ then A has more rows than columns and hence possesses a non-empty nullspace. In this case, given a $y \in Y$, in general there are many x's for which $y = Ax$. Hence A is not injective. If the rank of A is m however then it should not be hard to see that A is however surjective. Now if $m = n$ and the A has full rank, then this linear operator is bijective. In fact, it is a well know fact from linear algebra that such matrices possess an inverse. Hence bijectivity is basically the same as invertability.

Finally, to compute the norm of A, we here consider only the square case for which $\|Ax\|_2^2 = x^T A^T A x$. Hence from the definition of $\|A\|$ we have

$$\|A\| = \max_{\|x\|=1} x^T A^T A x,$$

but this is just a definition (specifically from the Rayleigh quotient) of the maximal eigenvalue of $A^T A$.

Example 2.18 Let us define the space of interest as $X = C[0, 1]$, the space of all continuous functions on the interval $[0, 1]$. The one example (really the canonical example) of a linear operator, A, on this space is (note the various notations)

$$y(s) = Ax = (Ax)(s) = \int_0^1 K(s, t) x(t) dt \tag{2.2}$$

The norm of A is

$$\|A\| = \max_{0 \le s \le 1} \int_0^1 |K(s, t)| dt$$

This example shows that linear operators in a "continuous" setting have essentially the same form as those in a discrete, matrix-vector setup. Specifically comparing (2.1) and (2.2) shows that the discrete sum is merely replaced by a continuous integral with the finite dimensional matrix A taking the form of a function of two variables, $K(s,t)$ (sometimes refered to as the *kernel* of the operator) with s the "row" variable and t the "column."

Example 2.19 A final class of linear operators which we shall encounter throughout this manuscript is an orthogonal projector. Given an arbitrary element $x \in X$ and a subspace F, the Projection Theorem basically guarantees that there is a closest point $f \in F$ to $x \in X$, but does not tell us how to find f. If X is a Hilbert space, F is finite dimensional, and we know an orthogonal basis for F, we can explicitly construct a linear operator that takes x to its closest point in F. Specifically, suppose the set of N mutually orthogonal vectors ϕ_n, $n = 1, 2, \ldots N$ span F then the orthogonal projector onto F is the linear operator $P : X \to F$ defined as

$$Pf = \sum_{i=1}^{N} (f|\phi_i)\phi_i \tag{2.3}$$

In the case where X and F are finite dimensional, we can gather together the orthogonal vectors spanning F into a matrix Φ whose i-th column is just ϕ_i. Then (2.3) takes the simpler form

$$\begin{aligned} Pf &= \sum_{i=1}^{N} (\phi_i|f)\phi_i = \begin{bmatrix}\phi_1 \mid \phi_2 \mid \ldots \mid \phi_N\end{bmatrix} \begin{bmatrix}\phi_1^T f \\ \phi_2^T f \\ \vdots \\ \phi_N^T f\end{bmatrix} \\ &= \begin{bmatrix}\phi_1 \mid \phi_2 \mid \ldots \mid \phi_N\end{bmatrix} \begin{bmatrix}\phi_1^T \\ \phi_2^T \\ \vdots \\ \phi_N^T\end{bmatrix} f = \Phi\Phi^T f \end{aligned}$$

and we can conclude that an explicit representation for P is the outer-product matrix $\Phi\Phi^T$.

In general, a projector P is any operator which satisfies $P^2 = P$ and $P = P^*$.[4]

Again thinking of the finite dimensional case, one of the most fundamental matrices related to A is its transpose, A^T obtained by "swapping" rows and columns, $A^T_{i,j} = A_{j,i}$. The generalization of transpose to infinite dimensional operators is provided by the *adjoint*.

Definition 2.28 *Say X and Y are normed linear vector spaces and A a linear transformation from X to Y. The* adjoint *of A, denoted as A^*, is a mapping from the dual space of Y to the dual space of X, $A^* : Y^* \to X^*$ defined according to*

$$\langle x, A^* y^* \rangle = \langle Ax, y^* \rangle \tag{2.4}$$

Moreover, it can be shown that $\|A^*\| = \|A\|$.

[4]EXERCISES: Verify these for (2.3) and do the case of nonorthogonal projectors.

This definition of the adjoint takes some getting used to and is best interpreted one piece at a time. Let us start by fixing some element in the dual space of Y, $y^* \in Y^*$. In this case $\langle Ax, y^* \rangle$ is a scalar function for $x \in X$. Hence, it must be a linear functional on X. In fact, it can be shown that this linear functional is bounded. Hence, it must correspond to some $x^* \in X^*$. That is when viewed as a function of X, $\langle Ax, y^* \rangle$ is an element of the dual space of X. That is there is some $x^* \in X^*$ for which $\langle Ax, y^* \rangle = \langle x, x^* \rangle$. We define this element to be $A^* y^*$.

In the case where X and Y are both Hilbert spaces, more can be said about the adjoint and in general the discussion simplifies considerably. First, we can make use of the inner product associated with these spaces to conclude that the adjoint satisfies $(Ax|y) = (x|A^*y)$. Second, if X and Y are Hilbert spaces, they are also self-dual so $A^* : Y \to X$. Third, the adjoint of the adjoint is the original operator, $A^{**} = A$. If $A = A^*$ then the operator is called *self-adjoint*. Finally, if A is self adjoint and $(x|Ax) \geq 0$ then the operator is called *positive semi-definite*. Thus self-adjoint operators represent generalizations of symmetric matrices and positive semi-definite operators generalize the idea of positive semi-definite matrices. Finally, given a linear operator A, determination of its adjoint is a fairly straightforward exercise as we now show on a couple of examples:

Example 2.20 In the finite dimensional case we have $X = \mathbb{R}^n$. $Y = \mathbb{R}^n$, and A an $m \times n$ matrix. To find the adjoint of A we manipulate $(Ax|y)$ to obtain an expression of the form $(x|$ something $y)$ That "something" must, by the definition be the adjoint. Mathematically we have

$$\begin{aligned} (Ax|y) &= \sum_{i=1}^{m} \sum_{j=1}^{n} y_i A_{i,j} x_j && (2.5) \\ &= \sum_{j=1}^{n} x_j \left[\sum_{i=1}^{m} A_{i,j} y_i \right] && (2.6) \\ &= (x|A^*y). && (2.7) \end{aligned}$$

Hence the adjoint in this case involves summing over "rows" rather than "columns" and we may identify $A^*_{i,j} = A_{j,i}$. Thus, in the finite dimensional case, A^* is just the transpose of A.

Example 2.21 For a slightly more interesting example, let $X = Y = L_2[0, 1]$ and

$$Ax = \int_0^t K(t, s) x(s) ds \tag{2.8}$$

Note the t in the limits of the integral. Following analogous steps to the matrix case we compute

$$(Ax|y) = \int_0^1 dt\, y(t) \left[\int_0^t ds\, K(t, s) x(s) \right]. \tag{2.9}$$

Pictorially, the integration in (2.9) is shown in the left hand side of Fig. 2.3. To accomplish the equivalent of summing over rows rather than columns as we did for the matrix case, we need to change the order of integration to that shown in the left side of Fig. 2.3. That is:

$$\begin{aligned} (Ax|y) &= \int_0^1 x(s) ds \left[\int_s^1 dt\, K(t, s) y(t) \right] && (2.10) \\ &= (x|A^*y). && (2.11) \end{aligned}$$

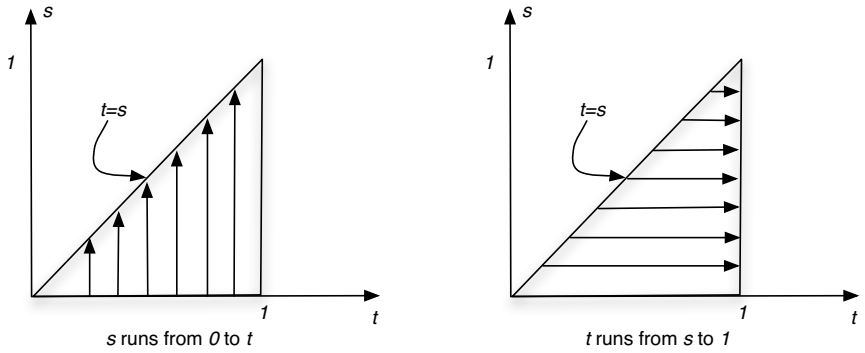


Figure 2.3: Reversing the order of integration to find the adjoint of the operator in (2.8)

That is, by carefully reversing the order of integration, we can express the original inner product with y as an inner product of x against something that looks like a linear opertator acting on y. That operator must be the adjoint. Hence we conclude:

$$(A^*y)(s) = \int_s^1 K(t,s)y(t)dt$$

Provided with the notion of an adjoint and in the case where X and Y are Hilbert spaces, the nullspace and range of an operator, A provide a very elegant decomposition of X and Y. As shown in Fig. 2.4, the input space can be decomposed into two orthogonal subspaces, the nullspace and the range of the adjoint of A. Similarly, Y admits an orthogonal decomposition into the range of A and the nullspace of A^* [5] Mathematicaly then we have the following relationships

$$\overline{\mathcal{R}(A)} = [\mathcal{N}(A^*)]^\perp \quad \text{and} \quad \overline{\mathcal{R}(A^*)} = [\mathcal{N}(A)]^\perp$$

so that we can write

$$X = \mathcal{N}(A) \oplus \overline{\mathcal{R}(A^*)} \quad \text{and} \quad Y = \mathcal{N}(A^*) \oplus \overline{\mathcal{R}(A)}$$

In finite dimensions, it is not hard to show for example that the nullspace of A is orthogonal to the range of A^*. Indeed, say that we have a vector $x_n \in \mathcal{N}(A)$. This means that $Ax_n = 0$. Next suppose that $x_r \in \mathcal{R}(A^*)$. This means there is a vector y for which $A^*y = A^Ty = x_r$. Now we have $x_r^T x_n = yA^T x_n = y\left(A^T x\right) = 0$ since x_n is, by assumption, in the nullspace of A. Hence a vector in $\mathcal{N}(A)$ is perpendicular to one in $\mathcal{R}(A^*)$.

2.2.3 Eigenanalyis and the Singular Value Decomposition

To generalize the ideas of eigenvectors and eigenvalues as discussed in an introductory treatment of linear algebra, let us briefly review the case of the convolution integral which arises in the study

[5] Technically, the overbars in Fig. 2.4 indicate that one must consider the closure of the range of A and its adjoint in defining these subspaces. Roughly speaking, the closure of a space is the space itself plus all limit points (which for whatever reason are not included in the definition of the space) associated with convergent sequences in that space. Intuitively, one can think of the closure as the interior of the space plus its "boundary." This may be visualized in 2 dimensions as e.g., the internal points of a circle (it's interior) plus the boundary of the circle (the limit points).

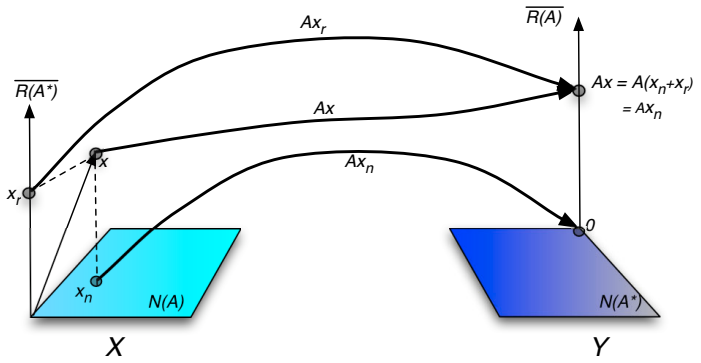


Figure 2.4: Four fundamental subspaces induced by an operator acting on two Hilbert spaces.

of linear systems theory. For a system that is both linear and time invariant, the output, $y(t)$ due to an input $x(t)$ may be computed as

$$y(t) = \int_{-\infty}^{\infty} h(t-s)x(s)\,ds \tag{2.12}$$

where $h(s)$ is known as the impulse response of the system; that is the output seen when the input is a Dirac delta function, $\delta(t)$.

It is well known that an alternate way of obtaining $y(t)$ is via the use of Fourier transform analysis. Here we take the Fourier and inverse Fourier transforms to be defined according to

$$X(\omega) = \mathcal{F}(x) = \int_{-\infty}^{\infty} x(t)e^{i\omega t}\,dt = (x(t)|e^{-i\omega t}) \tag{2.13}$$

$$x(t) = \mathcal{F}^{-1}(X) = \frac{1}{2\pi}\int_{-\infty}^{\infty} X(\omega)e^{-i\omega t}\,d\omega = (X(\omega)|e^{i\omega t}). \tag{2.14}$$

In other words, the Fourier transform of a signal x at the frequency ω is the inner product of that signal with the function $e^{i\omega t}$ with an analogous interpretation holding for the inverse Fourier transform.[6] With these definitions, we have

$$y(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} H(\omega)\left(x(t)|e^{-i\omega t}\right)e^{i\omega t}d\,\omega. \tag{2.15}$$

That is, the convolution operator may be evaluated in three steps:

1. Compute the Fourier transform of x by taking the inner product of x with the set of "basis" functions, $e^{i\omega t}$. Unlike our previous discussion, these basis functions are indexed by a continuous variable ω, rather than $i = 1, 2, \ldots$, but the basic idea is the same. The important insight here is to think of the Fourier transform as a mathematical prism that *analyzes* $x(t)$ into its constituent sinusoidal components.

[6]To define an inner product, we do of course require an underlying vector space. For simplicity we take this space to the $L_2[-\infty, \infty]$ and conveniently ignore all of the subtleties associated with issues like the convergence of the integrals, the existence of δ functions and such. These important details may be found in more advanced mathematics texts such as XXX.

2. Weight each of the Fourier coefficients by a factor $H(\omega)$. In the signal processing literature, this weighting function is known as a filter, a terminology we shall adopt in throughout this text. The resulting weighted input is $Y(\omega)$ the Fourier transform of the output signal, y.

3. Reassemble (synthesize) the resulting output signal $y(t)$ through the use of the inverse Fourier transform

Thus, the Fourier transform is said to turn the fairly complicated linear operator convolution into simple multiplication. In a sense, if one wants to compute a convolution, the "right" domain in which to work is not t, but rather ω where the operation is trivially done on a frequency-by-frequency basis. The reason this is true is that complex exponentials are the eigenfunctions of convolution. Without being too precise at this point, an eigenfunction of an operator A is a function that remains unchanged up to a complex scaling when passed through A. Mathematically, for ϕ and eigenfunction of A, we have $(A\phi)(t) = \lambda\phi(t)$ with $\lambda \in \mathbb{C}$. In the case of convolution with an impulse response $h(t)$, the eigenfunctions are of the form $e^{i\omega t}$ and the corresponding λ are $H(\omega)$ and clearly depend on ω.

In moving forward, it is useful to understand that this well-known example involving convolution and the use of Fourier transforms embodies just about all of the components of the far more general class of linear operators and the inverse problems associated with them:

1. We have a linear operator. In this case the kernel of the operator $K(t, s)$ is a function only of the difference of the arguments, $t - s$. In general this will not be the case.

2. We have a set of orthonormal eigenvectors (or eigenfunctions). For convolution they are the complex exponentials and are functions of a continuous variable, ω. For most of the problem we shall encounter, the eigenfunctions are indexed by the natural numbers, $n = 1, 2, 3, \ldots$.

3. For each eigenvector, we have an associated eigenvalue. For the convolution problem, the eigenvalues are generally complex numbers. For most of the remainder of this text, the quantities playing the role of the eigenvalues will be real values and non-negative. In any case, it is customary to call the collection of eigenvalues the *spectrum* of A whether or not the eigenvectors come from a Fourier basis..

4. These eigenvectors represent the natural basis in which to examine A. Specifically by expressing x in the eigenbasis of A, the action of A is simple multiplication one eigenvalue at a time.

5. Using the eigenstructure of A, evaluating Ax is a three step process: analysis-filtering-synthesis. For the convolution problem, analysis and synthesis are done using the same set of functions, the complex exponentials. As we move forward we shall see that two sets of orthonormal functions are required: one for analysis and a different set of synthesis.

Given this example, we can now discuss the eigenstructure of a linear operator in a slightly more formal manner [52, Section 15.3].

Definition 2.29 *For $A : X \to X$ a bounded linear operator and X a normed space, the scalar λ is said to be an* eigenvalue *of A and the function ϕ an* eigenfunction *if $A\phi = \lambda\phi$; that is if ϕ is in the nullspace of $A - \lambda I$ with I the identity operator on X.*

By imposing additional structure on A significantly more can be said concerning its eigenvalues and eigenfunctions. If A is self-adjoint and compact[7] then the following can be shown to hold:

1. The eigenvalues of A are all real.
2. At least one eigenvalue is not equal to zero.
3. The nullspace of $A - \lambda I$, $\mathcal{N}(A - \lambda I)$, are finite dimensional.
4. For $\lambda_i \neq \lambda_j$, $\mathcal{N}(A - \lambda_i I) \perp \mathcal{N}(A - \lambda_j I)$.
5. Say we order the eigenvalues so that $|\lambda_1| \geq |\lambda_2| \geq \ldots$ and define the operator $P_n : X \to \mathcal{N}(A - \lambda_n I)$ to be the orthogonal projector onto the nullspace of $A - \lambda_n I$. Then we can decompose A as
$$A = \sum_{n=1}^{\infty} \lambda_n P_n.$$
Thus, the action of A on a vector x can be expressed as
$$Af = \sum_{n=1}^{\infty} \lambda_n \sum_{j=1}^{N_j} (f|\phi_{n,j})\phi_{n,j} \tag{2.16}$$
where $\{\phi_{n,j}\}_{j=1}^{N_j}$ are an orthonormal set of vectors spanning $\mathcal{N}(A - \lambda_n I)$ and we have made use of the structure of an orthonormal projector defined on page 14.
6. If we take Q to be the orthonormal projector of onto $\mathcal{N}(A)$, then any $x \in X$ can be written as
$$x = Qx + \sum_{n=1}^{\infty} P_n x = \left[Q + \sum_{n=1}^{\infty} P_n \right] x.$$
That is we have a *resolution of the identity* in that we can write $I = Q + \sum_{n=1}^{\infty} P_n$. Moreover, we have deconposed the space X into a set of mutually orthogonal spaces [IS THIS TRUE WITH Q?] and thus can write $X = [\bigoplus_{n=1}^{\infty} \mathcal{N}(A - \lambda_n I)] \oplus \mathcal{N}(A)$

To gain an intutive understanding as to the import and utility of an eigendecomposition of an operator, let us consider a specific, finite dimensional case given by the matrix
$$A = \begin{bmatrix} 1.4698 & -0.5223 & -0.1634 \\ -0.5223 & 1.4848 & -0.1297 \\ -0.1634 & -0.1297 & 0.0453 \end{bmatrix}. \tag{2.17}$$
By direct calculation we see that A has three eigenvalues, $\{2, 1, 0\}$ and three orthonormal eigenvectors that can be arranged as columns in a matrix, Φ, which to four significant digits is
$$\Phi = [\phi_1 \mid \phi_2 \mid \phi_3] = \begin{bmatrix} -0.7036 & -0.6926 & 0.1588 \\ 0.7105 & -0.6894 & 0.1412 \\ 0.0117 & 0.2122 & 0.9772 \end{bmatrix}$$

[7] A compact operator is one for which all convergent sequences $\phi_n \in X$ generate continuous, convergent subsequences $A\phi_n$[CHECK THIS]

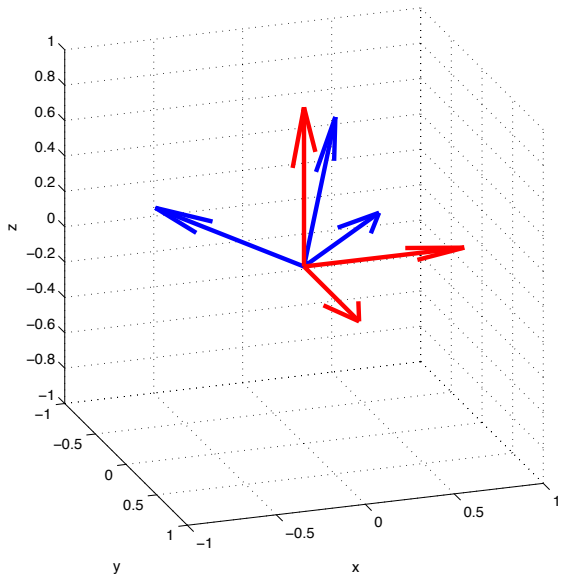


Figure 2.5: Eigen-structure of A in (2.17). The red vectors are the standard basis in $\mathbb{R}^3$. The blue vectors are the orthonormal vectors ϕ_1, ϕ_2 and ϕ_3 representing the eigenbasis for A.

Thus, we can write $A = \Phi\Lambda\Phi^T$ with Λ the diagonal matrix formed from the three eigenvalues. The orthonormality of Φ implies that

$$\Lambda = \Phi^T A \Phi \tag{2.18}$$

That is the matrix of eigenvectors *diagonalizes* A. Because $\Phi^T = \Phi^{-1}$, the right hand side of (2.18) is a similarity transformation of A; essentially a representation of A in a basis whose coordinate vectors are given by the eigenvectors. To see why such a representation is, in a sense, the "natural" one for A, consider the action of A on an arbitrary vector x:

$$Ax = \Phi\Lambda\Phi^T x \tag{2.19}$$

$$= \sum_{i=1}^{3} \phi_i \lambda_i \left(\phi_i^T x\right) \tag{2.20}$$

where (2.20) is equivalent to (2.15) in the convolution case or (2.16) in the general case. Equation (2.20) says that we can apply A to x using the same three steps we followed for the convolution-Fourier example:

1. **Analysis:** Compute $\phi_i^T x = (\phi_i | x) = \chi_i$. This operation can be interpreted in a number of equivalent ways. Each inner product is basically the projection of x onto each of the three natural axes used in representing A; *i.e.* the blue vectors in Fig. 2.5. Thus the scalar $\phi_i^T x$ is the "amount" of ϕ_i in x. Alternatively, by the orthonormality of Φ, we have that $\|\chi\|_2^2 = x^T\Phi\Phi^T x = x^T x = \|x\|_2^2$. That is, multiplication of x by Φ^T leaves the length of the vector unchanged. A linear transformation possessing this property is basically a rotation. In this case the vector χ is a rotation of x into the coordinate system whose axes are given by the eigenvectors of A.

2. **Filtering:** Compute the vector $\chi_1 = \Lambda\chi$. That is $\chi_{1,i} = \lambda_i\chi_i$. In other words the action of A on a vector already in the eigencoordinate system of A is simply scaling along each of the coordinate axes. Hence in the Φ basis, multiplication by the matrix A is replaced by the simpler operation of three scalar multiplications.

3. **Synthesis:** Compute $\Phi\chi_1$ which is just the rotation from the Φ coordinate system back to the original.

Finding an eigendecomposition for an arbitrary linear operator is no small task and indeed is the subject in its own right of many texts and much research. Here we demonstrate through an example on technique that can be used: turning a linear integral equation into a differential equation whose solution can be determined by standard methods. Toward this end, consider the operator [52, Example 15.13]

$$(Af)(x) = g(x) = \int_0^\pi K(x,y)f(y)\,dy \tag{2.21}$$

with

$$K(x,y) = \begin{cases} K_1(x,y) = \frac{1}{\pi}(\pi - x)y & 0 \le y \le x \le \pi \\ K_2(x,y) = \frac{1}{\pi}(\pi - y)x & 0 \le x \le y \le \pi \end{cases} \tag{2.22}$$

so that

$$g(x) = \int_0^x K_1(x,y)f(y)\,dy + \int_x^\pi K_2(x,y)f(y)\,dy. \tag{2.23}$$

We claim that g satisfies the boundary value problem

$$\frac{d^2}{dx^2}g(x) = -f(x) \quad g(0) = g(\pi) = 0. \tag{2.24}$$

This is shown by elementary calculus. Differentiating (2.23) once with respect to x and recalling Laplace's rule for differentiation of the limit of an integral yields

$$\frac{d}{dx}g(x) = \int_0^x \frac{d}{dx}K_1(x,y)f(y)\,dy + \int_x^\pi \frac{d}{dx}K_2(x,y)f(y)\,dy + K_1(x,x)f(x) - K_2(x,x)f(x)$$

but $K_1(x,x) = K_2(x,x) = 0$. Next, since $\frac{d}{dx}K_1(x,y) = -y/\pi$ and $\frac{d}{dx}K_2(x,y) = 1 - y/\pi$ a bit of algebra and one more derivative yield

$$\begin{aligned} \frac{d^2}{dx^2}g(x) &= \frac{d}{dx}\left[\int_0^x -\frac{y}{\pi}f(y)\,dy + \int_x^\pi \left(1 - \frac{y}{\pi}\right)f(y)\,dy\right] \\ &= -\frac{x}{\pi}f(x) - \left(1 - \frac{x}{\pi}\right)f(x) = -f(x) \end{aligned}$$

The boundary condition follows by substitution of $x = 0$ and $x = \pi$ directly into (2.23) and using the definitions of K_1 and K_2.

Now let us return to the eigenproblem $A\phi = \lambda\phi$ with A defined in (2.21). Differentiating twice and using (2.24) yields

$$\begin{aligned} \frac{d^2}{dx^2}\left[A\phi\right] &= -\phi \\ &= \lambda\frac{d^2}{dx^2}\phi \end{aligned}$$

with the boundary conditions $\phi(0) = \phi(\pi) = 0$. It is readily verified that a unit norm solution to this boundary value problem is $\phi_n(x) = \sqrt{\frac{2}{\pi}}\sin nx$ for $n = 1, 2, \ldots$ and $0 \le x \le \pi$ assuming $\lambda_n = \frac{1}{n^2}$ and hence these are the eigenfunctions and eigenvalues of (2.21). Moreover, these eigenfunctions are mutually orthogonal, $(\phi_n|\phi_m) = 0$ for $m \neq n$. Hence each of the P_n subspaces is spanned by a single eigenfunction and we can write

$$(Af)(x) = \sum_{n=1}^{\infty}\frac{1}{n^2}(f|\phi_n)\phi_n(x) \tag{2.25}$$

$$= \frac{2}{\pi}\sum_{n=1}^{\infty}\frac{1}{n^2}\sin nx\int_0^{\pi} f(y)\sin ny\,dy = \int_0^{\pi} f(y)\left[\sum_{n=1}^{\infty}\frac{2}{n^2\pi}\sin nx\sin ny\right]dy \tag{2.26}$$

and we conclude that the eigen-decomposition of K is

$$K(x,y) = \sum_{n=1}^{\infty}\frac{2}{n^2\pi}\sin nx\sin ny$$

To obtain an analysis-filtering-synthesis interpertation for the vast majority of operators that are not self adjoint, we move from the notion of an eigenvector-eigenvalue decomposition to the *singular value decomposition* (SVD). The SVD is like an eigendecomposition for the square of A. With operators though there are two natural ways of obtaining a square, namely AA^* and A^*A. In general, these two are not the same (if they are, A is said to *commute* with A^*[8]), but they are both self-adjoint and fit into the eigenanalysis framework we have just developed. Thus, it turns out that there are two sets of singular vectors in an SVD, one for each way of squaring A and remarkably, a single set of singular values. More formally we have the following [52, Section 15.4]

If $K(x,y)$ is square integrable[9] and we let $(Af)(x) = \int K(x,y)f(y)\,dy$ then the operators $L = A^*A$ and $L^* = AA^*$ are compact and positive semi-definite and A^*A and AA^* have the same set of eigenvalues, σ_k^2. If we further take u_k and v_k as the solution to the eigenproblems

$$Lu_k = \sigma_k^2 u_k \tag{2.27}$$

$$L^*v_k = \sigma_k^2 v_k \tag{2.28}$$

then

1. The set u_k are an orthonormal basis for $[\mathcal{N}(A)]^{\perp}$; that is, the set of f's not put to zero by A.

[8] Must check if this is an "if and only if" for self-adjoint?

[9] Need to define square integrable kernels

2. The set v_k are an orthonormal basis for $[\mathcal{N}(A^*)]^\perp$; that is, the closure of the range of A.

3. The action of A on a vector f can be written as

$$(Af)(x) = \sum_{k=1}^{\infty} \sigma_k (f|u_k) v_k(x) \tag{2.29}$$

and we define

Definition 2.30 *The* singular value decomposition *(SVD) of an operator A with a square integrable kernel $K(x, y)$ is composed of*

1. *The set of orthonormal vectors u_k called the* right singular vectors *satisfying (2.27)*
2. *The set of orthonormal vectors v_k called the* left singular vectors *satisfying (2.28).*
3. *The non-negative square roots of the eigenvalues of L and L^*, σ_k, called the* singular values *of A.*

Generally the SVD is computed such that the singular values are ordered $\sigma_1 > \sigma_2 > \sigma_3 > ... > 0$.

Equation (2.29) is just what we are looking for in terms of a convenient way of describing how A works. First there is an analysis step using the basis formed from the u_k. For this reason the quantities $(f|u_k)$ are often called generalized Fourier coefficients. Next the singular values are used to filter each generalized Fourier coefficient. Finally, the v_k are used to synthesize the output. The structure of (2.29) also makes clear the relationship of the singular vectors to the range and nullspace of A. If for all k, $f \perp u_k$, then clearly $Af = 0$. Hence the u_k span the space of vector orthogonal to the nullspace of A. Similarly, since every Af is composed of a linear combination of the v_k and the v_k are orthonormal, they must span the range of A.

Following the path we have taken before, to gain a more intuitive understanding of the SVD we examine in some detail its structure for finite dimensional operators, that is matrices. Thus, let A be a matrix with m rows and n columns. In this case, the SVD of A is of the form

$$A = V \Sigma U^T = \begin{bmatrix} V_1 \mid V_2 \end{bmatrix} \begin{bmatrix} \Sigma_1 & 0_{k,n-k} \\ 0_{m-k,k} & 0_{m-k,n-k} \end{bmatrix} \begin{bmatrix} U_1^T \\ U_2^T \end{bmatrix} \tag{2.30}$$

with $U^T U = I_n$, $V^T V = I_n$, $0_{k,l}$ the $k \times l$ matrix of zeros, I_n the $n \times n$ identity matrix and

$$\Sigma_1 = \begin{bmatrix} \sigma_1 & 0 & 0 & \dots & 0 \\ 0 & \sigma_2 & 0 & \dots & 0 \\ 0 & 0 & \sigma_3 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & 0 & \sigma_k \end{bmatrix}.$$

The matrices associated with U and V are sized as follows: $U : n \times n$, $U_1 : n \times k$, $U_2 : n \times n-k$, $V : m \times m$, $V_1 : m \times k$, and $V_2 : m \times m-k$. Thus for $f \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$ we have:

$$y = Af = \sum_{i=1}^{k} \sigma_i \left(u_i^T f\right) v_i \tag{2.31}$$

where u_i is the i-th column of the matrix U and similarly for v_i.

From this way of writing A, we obtain a very convenient way of characterizing the four fundamental subspaces. The nullspace of A is comprised of all those vectors for which $Af = 0$. If f is in the linear span of the k vectors of U_1, then, since the $\sigma_i > 0$, $Af \neq 0$. Hence the nullspace of A is spanned by the columns of U_2. Because U is orthonormal, the columns of U_1 must be a basis for $[\mathcal{N}(A)]^{\perp}$. Similar statements can be made about V and the range space of A. [EXERCISE].

2.3 Exercises

2.1 Here we examine some interesting gradient calculations. To start, we define the following:

- The derivative of a scalar function y with respect to a vector $\mathbf{x} \in R^n$ is the column vector, $\partial y/\partial \mathbf{x}$ whose ith entry is $\partial y/\partial \mathbf{x}_i$ where $\mathbf{x}_i$ is the ith element of the vector $\mathbf{x}$.
- The derivative of a vector, $\mathbf{y} \in R^m$ with respect to a vector $\mathbf{x} \in R^n$ as the matrix, $\partial \mathbf{y}/\partial \mathbf{x}$ whose entry on row i, column j is $\partial \mathbf{y}_j/\partial \mathbf{x}_i$.

Given these definitions prove the following

(a) Chain rule: If $\mathbf{z} = \mathbf{y}(\mathbf{x}) \in R^r$ then $\partial \mathbf{z}/\partial \mathbf{x} = (\partial \mathbf{y}/\partial \mathbf{x})(\partial \mathbf{z}/\partial \mathbf{y})$.

(b) $\frac{\partial}{\partial \mathbf{x}} \mathbf{A}\mathbf{x} = \mathbf{A}^T$

(c) $\frac{\partial}{\partial \mathbf{x}} \mathbf{x}^T \mathbf{A}\mathbf{x} = \mathbf{A}\mathbf{x} + \mathbf{A}^T\mathbf{x}$. Specialize to the case where $\mathbf{A}$ is symmetric.

(d) Let the invertible matrix $\mathbf{A}$ be a function of a scalar variable, x. Show that $\partial \mathbf{A}^{-1}/\partial x = -\mathbf{A}^{-1}\left(\partial \mathbf{A}/\partial x\right)\mathbf{A}^{-1}$. Hint: chain rule $\mathbf{A}(x)\mathbf{A}^{-1}(x) = I$.

(e) Letting $\mathbf{X}_{r,c}$ be the element of the matrix $\mathbf{X}$ on row r and column c show that $\partial\left(\mathbf{A}\mathbf{X}^{-1}\mathbf{B}\right)/\partial \mathbf{X}_{r,s} = -\mathbf{A}\mathbf{X}^{-1}\mathbf{E}_{r,c}\mathbf{X}^{-1}\mathbf{B}$ where $\mathbf{E}_{r,c}$ is the matrix of all zeroes except for a 1 in row r and column c.

2.2 Here we want to look at the least squares problem in a Hilbert space. Say that $\{y_i\}_{i=1}^n$ generate a closed, finite dimensional subspace M of a Hilbert space H.[10] For an arbitrary $x \in H$ we want to find that vector $x_m \in M$ which is closest to x in that $\|x - x_m\|$ is minimized.

1. Argue that the desire to minimize $\|x - x_m\|$ is equivalent to finding a collection of scalars $\{a_n\}_{i=1}^n$ which minimize
$$\left\| x - \sum_{i=1}^{n} a_i y_i \right\| \tag{2.32}$$
2. Let a be the vector of the a_i coefficients. Using the projection theorem, show that the a must satisfy a matrix-vector problem of the form
$$Ga = b \tag{2.33}$$
where G is symmetric and is termed the *Gram* matrix. What is the $ij$the entry of G and what are the elements of b?

[10] Recall that a subspace *generated* by a set of vectors is nothing more than all vectors which can be written as a linear combination of the set in question. It is not necessarily the case that this set is a basis.

3. Prove that G is invertible if and only if the vectors y_i are independent.

2.3 Adjoints of operators appear throughout reconstruction and inverse theory. In this problem we will study such adjoints operators. Recall, given a linear operator $L : \mathcal{X} \mapsto \mathcal{Y}$, the adjoint L^* is defined by the relationship:

$$\langle y, Lx \rangle_{\mathcal{Y}} = \langle L^* y, x \rangle_{\mathcal{X}} \tag{2.34}$$

for all x, y, where $\langle \cdot, \cdot \rangle_{\mathcal{X}}$ denotes the inner product in the space $\mathcal{X}$ and $\langle y, Lx \rangle_{\mathcal{Y}}$ denotes the inner product in the space $\mathcal{Y}$.

(a) Suppose the linear operator is represented by a real integral equation of the first kind of the form:

$$Lx(t) = \int K(t, \tau) x(\tau) \, d\tau \tag{2.35}$$

and we take the usual inner product between two real functions $u(t)$, $v(t)$:

$$\langle u, v \rangle = \int u(t) \, v(t) \, dt \tag{2.36}$$

What is the corresponding adjoint operator in this case? An operator is termed "self-adjoint" when it equals its adjoint. What property must the kernel satisfy for the linear operator to be self-adjoint?

(b) Next consider the case of LTI filtering or convolution:

$$Lx(t) = \int h(t - \tau) x(\tau) \, d\tau \tag{2.37}$$

What is the adjoint operator to convolution? What condition does the impulse response have to satisfy for corresponding convolution operation to be self-adjoint?

2.4 For integral equations where the observed quantity can be viewed as the output of an LTI system one is often lead to consider discrete representations with a corresponding convolutional structure. In this problem we investigate such discrete LTI inverse problems, i.e. problems of the form:

$$y(i) = \sum_{j=1}^{L} h(i - j) x(j) = h * x \tag{2.38}$$

where the nonzero portion of $h(i)$ is of length P and that of $x(i)$ is of length L and $*$ denotes linear convolution.

(a) If we let y be the vector of $y(i)$ elements and x be the corresponding vector of $x(i)$ elements, what is the matrix C relating y and x through linear convolution $y = Cx$? What special form does it have? Note that the MATLAB function **convmtx** will generate the linear convolution matrix C for a given $h(i)$ and problem size.

(b) Since (2.38) represents a convolution we know that Fourier techniques should be useful. In particular, since the problem is discrete the appropriate tool is the discrete Fourier transform (DFT), which may be efficiently found using the FFT algorithm. But recall that the product of the DFT coefficients of two sequences actually corresponds to the *circular* convolution of the two sequences. In general, what length N circular convolution must be used to ensure that the circular convolution of $h(i)$ and $x(i)$ produce the same results as the linear convolution of these sequences? How are the corresponding periodic sequences $\widetilde{h}$ and $\widetilde{x}$ related to h and x? Write a MATLAB routine **cconv.m** to perform the N-point circular convolution of two sequences (Hint: Use the MATLAB **fft** routine).

(c) Given a sequence $h(i)$, what is the matrix $\widetilde{C}$ that performs the N-point circular convolution of this sequence with a length N vector (assuming $N > P$)? What special form does it have? How is it related to C? Using your routine **cconv.m**, write a MATLAB function **cconvmtx.m** to create $\widetilde{C}$ for an arbitrary h and N (Hint: Consider the relationship between circular convolution with the unit coordinate vectors and the columns of $\widetilde{C}$). Note that $y = \widetilde{C}\widetilde{x}$.

(d) The DFT pair of a discrete sequence $x(n)$ of length N is most commonly defined as:

$$\begin{aligned} X(k) &= \sum_{n=0}^{N-1} x(n)e^{-j2\pi kn/N} \qquad k = 1, \ldots, N \\ x(n) &= \frac{1}{N}\sum_{k=0}^{N-1} X(k)e^{j2\pi kn/N} \qquad n = 1, \ldots, N \end{aligned}$$

As usual, let X denote the vector of coefficients $X(k)$ and x denote the vector of the points $x(n)$. Then X and x are related by a matrix F through $X = Fx$, i.e. the matrix F takes the DFT of a sequence. The MATLAB function **fft** generates the N point DFT X of x. Using this function, write another MATLAB function **dftmtx.m** to generate the matrix F for an arbitrary N (Hint: Use the same approach as in part (c) and relate the DFT of the unit coordinate vectors to the columns of F). Write F^{-1} (the inverse DFT matrix) in terms of F itself (Hint: Note that two different columns of F are orthogonal).

(e) Let $Y = Fy$, $\widetilde{H} = F\widetilde{h}$ and $\widetilde{X} = F\widetilde{x}$ be vectors of DFT coefficients. Using the fact that $y = Cx = \widetilde{C}\widetilde{x}$, where $\widetilde{C}$ represents a circular convolution, together with the relationship between the DFT and circular convolution, show that the matrix $\widetilde{\mathbf{C}} = F\widetilde{C}F^H/N$ is a diagonal matrix, where F^H denotes the complex conjugate transpose of F. It is a general fact that circulant matrices are diagonalized by the DFT. In terms of operations on the rows and columns of $\widetilde{C}$ what does $F\widetilde{C}F^H/N$ represent? Verify these relationships for a numerical example, i.e. show that $y = \widetilde{C}\widetilde{x}$, $Fy = (F\widetilde{C}F^H/N)F\widetilde{x} = \widetilde{H}.*\widetilde{X}$, and that $F\widetilde{C}F^H/N$ is indeed diagonal.

(f) What is the relationship between the elements on the diagonal of $\widetilde{\mathbf{C}}$, the DFT of the first column of $\widetilde{C}$, and the DTFT of the original impulse response h? How are the diagonal elements of $\widetilde{\mathbf{C}}$ related to the eigendecomposition and the singular values of the circular convolutional matrix $\widetilde{C}$. What must the eigenfunctions of $\widetilde{C}$ be? How is the conditioning of the periodic system related to $\widetilde{\mathbf{C}}$? What is the physical interpretation of this result?

2.5 A large number of inverse problems require the solution to a linear least squares problem of the form:

$$\hat{f} = \arg\min_f \|y - Kf\|_2^2 + \sum_{i=1}^{N_L} \lambda_i^2 \|L_i f\|_2^2 \tag{2.39}$$

where K is a generally non-square matrix of size $M \times N$, y and f are appropriately sized vectors, λ_i are scalars, and L_i are a $P_i \times N$ matrices.

1. Show that the solution to (2.39) is equivalent to solving

$$\hat{f} = \arg\min_f \|y_1 - K_1 f\|_2^2 \tag{2.40}$$

for some vector y_1 and matrix K_1.

2. Show that $\hat{f}$ is in fact the solution to a linear system of equations defined explicitly by y, K, L_i, and λ_i. Find that system and comment on the conditions for that linear system to have a unique solution.

2.6 Let $v_1, \ldots, v_k$ be a basis for a finite-dimensional space, $\mathcal{V}$. It follows from the definition of a basis that any vector $v \in \mathcal{V}$ can be written in the form: $\sum_{i=1}^k \alpha_i v_i$ for some set of α_i. Show that the choice of α_i in this representation is unique.

2.7 (a) Let M be a closed subspace of a Hilbert space H. The operator P (called the projection operator onto M) defined by $Px = x_m$ where $x = x_m + x_n$ is the unique representation of $x \in H$ with $x_m \in M$ and $x_n \in M^\perp$. Show that the projection operator is linear and has norm equal to 1.

(b) Show that a bounded linear operator on a Hilbert space H is a projection operator if and only if $P^2 = P$ (indempotent) and $P^* = P$ (self-adjoint).

(c) Two projection operators P_1 and P_2 on a Hilbert space are said to be orthogonal if $P_1 P_2 = 0$. Show that two projection operators are orthogonal if and only if their ranges are orthogonal

2.8 Let M and N be orthogonal closed subspaces of a Hilbert space H and let x be an arbitrary vector in H. Show that the subspace $M \oplus N$ is closed and that the orthogonal projection of x onto $M \oplus N$ is equal to $x_m + x_n$ where x_m is the orthogonal projection of x onto M and x_n is the orthogonal projection of x onto N.

Chapter 3

A Collection of Forward and Inverse Problems

This chapter is concerned with the development of the classes of forward models and associated inverse problems that will drive most all of the inverse methods we describe subsequently. As discussed in Chapter 1, four classes of problems are of interest: deconvolution, X-ray tomography, and inverse source and inverse scattering problems encountered when dealing with a scalar Helmholtz type of equation. For deconvolution and inverse scattering, the physics of the problem can be captured equally well using the tools of differential equations (ordinary as well as partial) or integral equations. Both formulations are presented here. Linear inverse problems are most naturally suited to an integral formulation. In the nonlinear case, there are advantages and drawbacks to either method as we discuss later. Unlike deconvolution and inverse scattering, tomography and the inverse source problems both are most easily modeled using integral equations.

3.1 Deconvolution

Convolution is perhaps the first nontrivial linear integral operator encountered at the undergraduate level in the study of ordinary differential equations or linear systems theory. It arises when one wishes to determine the solution to a linear constant coefficient differential or equation with zero initial conditions. While multi-dimensional forms of deconvolution are certainly of interest in a wide range of applications, here we restrict ourselves to the 1D, "temporal" case. Instances of the inverse source problem will be used to illustrate three dimensional deconvolution problems later in this chapter. The image processing literature provides copious examples of image restoration problems which basically are two dimensional deconvolution problems.

As indicated in the last paragraph, the modeling structure giving rise to a convolution is an N-th order ordinary constant coefficient differential equation:

$$\left[1 + \sum_{n=1}^{N} a_n \frac{d^n}{dt^n}\right] g(t) = \left[\sum_{n=0}^{M} b_n \frac{d^n}{dt^n}\right] f(t) \tag{3.1}$$

with initial conditions that g and its first $N-1$ derivatives are all zero. Taking the Fourier transform

of both sides of (3.1) and rearranging gives $G(\omega) = H(\omega)F(\omega)$. As discussed in § 2.2.3, this implies

$$g(t) = \int h(t-s)f(s)\,ds \tag{3.2}$$

where $h(t)$ is the inverse Fourier transform of the frequency response of the system

$$H(\omega) = \frac{\sum_{n=0}^{M} b_n(i\omega)^n}{1 + \sum_{n=1}^{N} a_n(i\omega)^n} \tag{3.3}$$

Taking the input to the system to be a Dirac delta function, $f(t) = \delta(t)$, we see from (3.2) that $g(t) = h(t)$. Thus the function $h(t)$ is also said to be the *impulse response* of the system. The linearity of (3.1) and the fact that the a and b coefficients are independent of time can be used to show that the impulse response is *all* we need to completely characterize the behavior of such a system with zero initial conditions.

Another way of interpreting the impulse response that will be of use later is as a "Green's function." To make the link clearest, consider the special case where $M = 0$ and $b_0 = 1$. In this case the problem of computing g from f may be written using operator theoretic notation as:

$$(Dg)(t) = f(t) \quad D = \sum_{n=0}^{N} a_n \frac{d^n}{dt^n} \tag{3.4}$$

plus the initial conditions. But we know that the solution to this problems is given by

$$g(t) = (Af)(t) = \int h(t-s)f(s)\,ds \tag{3.5}$$

where h is the inverse Fourier transform of $H(\omega) = \left[1 + \sum_{n=1}^{N} a_n(i\omega)^n\right]^{-1}$. Thus, symbolically at least we conclude that

$$(ADg)(t) = g(t) = (Af)(t) \tag{3.6}$$

so that A is a left inverse of D. That is, the inverse of the differential operator D (plus initial conditions) is the integral operator A. The kernel of this operator is called the *Green's function* for the problem. More specifically, the Green's function is defined to be the response of the system at time t to an impulse source at time s; that is as solution to $(Dh)(t,s) = \delta(t-s)$. By the time invariance of the problem though $h(t,s) = h(t-s)$.

To conclude, the discussion in this section has been concerned with convolution, the calculation of the output of a linear time invariant system to an arbitrary input. The deconvolution problem is basically the inverse of this: the recovery of the input, $f(t)$ given knowledge of $g(t)$ and $h(t)$ for all time. To zeroth order, the solution is quite straightforward. If we pass g through a system whose impulse response is the inverse Fourier transform of $H^{-1}(\omega)$, simple Fourier analysis shows that we should recover f. While this may be true in theory as we shall see in § 5.2, difficulties arise both in the event that $\|H(\omega)\| = 0$ for any ω, and more subtly, when $\|H(\omega)\|$ is "small" compared say to the amplitude of noise in the system at frequency ω.

3.2 X-ray Tomography

X-ray tomography represents perhaps the most basic inverse problem where we must determine the internal structure of a medium based on data obtained externally. The application of this method is perhaps best known in the context of medical imaging and various forms of nondestructive test and evaluation where it forms the basis for Computer Axial Tomography (CAT). The mathematical model for X-ray tomography that we develop here though is descriptive of a far broader class of sensing modalities than just CAT. Indeed, the analytics underlying the other medical imaging mainstay, Magnetic Resonance Imaging (MRI) are, for entirely different reasons, identical to those of CAT as is also the case for many newer imaging modalities such as Positron Emission Tomography (PET) and Single Photon Emission Computed Tomography (SPECT). Finally, under certain simplifying assumptions, the use of synthetic aperture techniques in radar signal processing also admit an X-ray tomographic-type of model.

To a very good approximation, X-rays travel through any reasonable medium in a straight line. To measure the rays after they pass through the region of interest then, we place a detector directly on across from a source of X-rays as shown in Fig. 3.1. While they do not scatter from their path, the effect of the material is to attenuate the intensity of the X-rays. The fractional attenuation in the beam as it passed though an infinitesimal length of the path ds is proportional to the length along the path multiplied by the density of the material, $f(x, y)$. Mathematically this relationship takes the form

$$\frac{\Delta I}{I} = -f(x, y)ds. \tag{3.7}$$

Assuming an intensity I_S at the source, (3.7) can be integrated to give the intensity at the detector, I_D as

$$I_D = I_S \exp\left\{\int_{ray} f(x, y)\, ds\right\} \Rightarrow \ln \frac{I_S}{I_D} = \int_{ray} f(x, y)\, ds. \tag{3.8}$$

Thus, X-ray tomography refers to the problem of recovering $f(x, y)$ from the integral of this quantity along lines.

Clearly, many such ray integrals will be required to have any hope of determining $f(x, y)$. Depending on the application and the instrumentation, a number of options exist for collecting tomographic information including parellel beam, fan-beam, helical beam and cone beam [47]. The simplest case to analyze is the parallel beam case which is shown in Fig. 3.1. Here a line of detectors is arranged across from a line of sources. The configuration is rotated around the object. Hence the data are collected as a function of θ, the angle made by the sources and detectors with the x axis, and t, the length along the detector array.

To arrive at a final model requires the analytical specification of the lines over which the integration takes place as functions of t and θ. As shown in Fig. 3.2, let $p = [x\ y]^T$ be a point in 2D and $u_\theta = [\cos\theta\ \sin\theta]^T$ be a unit vector in the direction θ. Geometrically the quantity $u_\theta^T p$ is the projection of p in the direction θ. Now, a point is on the $t - \theta$ ray if this projection is in fact equal to t. That is, the line of integration is the locus of points in the plane for which $t = x\cos\theta + y\sin\theta$ which we write formally as $\delta(t - x\cos\theta - y\sin\theta)$. The final X-ray tomographic model is

$$g(t, \theta) = \int_{-\infty}^{\infty} f(x, y)\delta(t - x\cos\theta - y\sin\theta)\, dx\, dy. \tag{3.9}$$

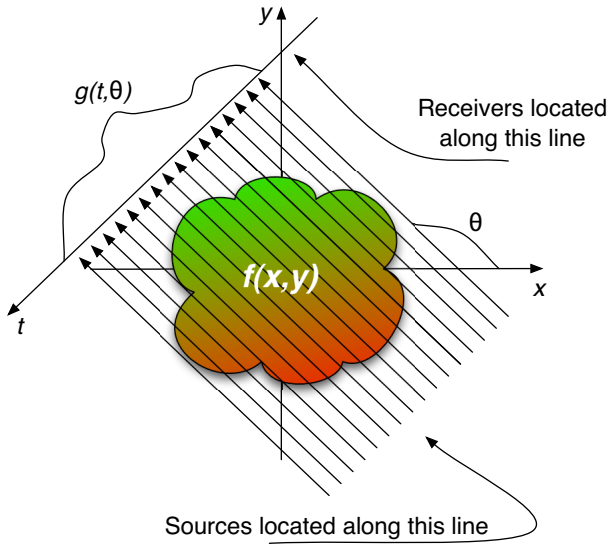


Figure 3.1: Parallel beam X-ray tomography

This mapping from $f(x, y)$ to $g(t, \theta)$ is known as the *Radon transform* of f. The transform itself as well as a wide range of generalizations has received considerable attention in the pure and applied mathematics, physics, and engineering communities. The interested reader is referred to [47,69] for additional information. In this text, we are only concerned with the most basic form of the Radon transform as given in (3.9).

3.3 Inverse Source and Inverse Scattering Problems

3.3.1 The Helmholtz Model

A key physical characteristic of the X-ray tomography problem is the assumption that the rays of energy propagate in straight lines through the medium only undergoing attenuation. In many application areas, it is necessary to account for a wider range of mechanisms by which energy can interact with the host medium. Most notable among these complicating factors is the process of scattering which essentially causes the rays to deviate from straight lines. To account for scattering effects, forward models more complex than the Radon transform are required. Here there are many options of vastly varying complexities depending on the sensor and its intended use. For problems involving electromagnetic sensing technologies (DC fields, eddy currents, radar, and even optical), the exact model is provided by the Maxwell's equations and requires detailed description of the space-time variations of three electrical and three magnetic field components. Similarly, the use of mechanical vibrations (e.g. acoutics) to excite a medium yields an exceptionally complex model governed by the laws of elasticity. Both the Maxwell's equations and elasticity in their most general forms are far to complex to be of any practical use in an introductory study of inverse problems. Indeed in both cases, many issues of forward modeling are current active areas of research.

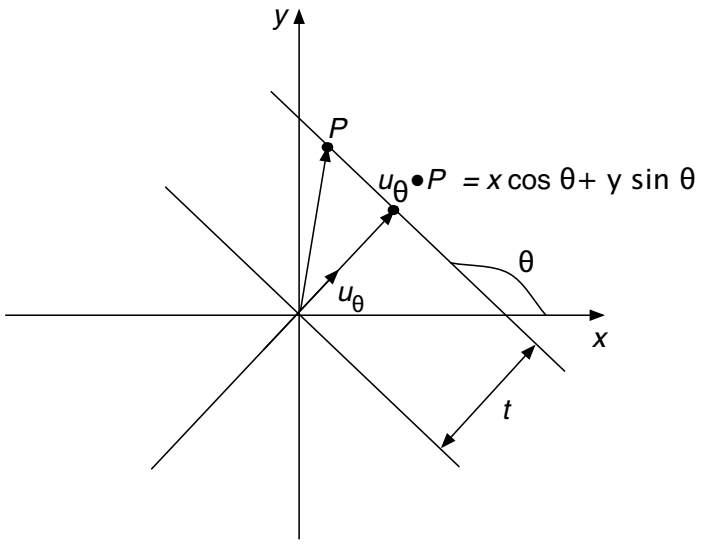


Figure 3.2: The geometry of parallel beam tomography

Luckily however, under certain reasonable simplifying assumptions, many inverse problems for electromagnetic, acoustic as well as other sensors can be reasonably modeling using a scalar Helmholtz-type of equation or its zero frequency limit, Poisson's equation. In addition to the applicability of this model, it also represents a useful first step in moving beyond the Radon transform for two reasons. First, the Helmholtz model provides a reasonably tractable example of the inherently nonlinear nature of inverse problems whose phenomenology is governed by processes more complex than attenuation. Second, powerful methods of linearizing these inverse problems are easily illustrated within the context of the Helmholtz equation.

The scalar Helmholtz equation of interest here is a partial differential equation governing the spatial distribution of a field, $\phi(\mathbf{r})$. In all cases of practical interest, this model is obtained as a result of the use of a time harmonic excitation, $e^{i\omega t}$, for a physical problem whose temporal behavior is governed by a wave equation. Thus, ϕ (as well as other quantities in the model) are also functions of the temporal frequency ω; although for notational simplicity, this dependence is not usually made explicit. Two classes of problems are of interest: those for which the domain is closed and those for which it is open. Closed problems involve fields in some finite region of space, Ω, which is bounded by a collection of boundaries, $\partial\Omega$. Open problems have no such boundaries but rather require that the field be computed over all points in space. In both cases, the basic governing equation is the same. Analogous to the requirement of specifying initial conditions for (3.1), for closed problems we must specify a set of boundary conditions the field or its gradient must satisfy on $\partial\Omega$. For open problems it turns out that the asymptotic requirements on the behavior of the field as $|r| \to \infty$ serve the same purpose as boundary conditions.

The most general governing equation we shall use here is:

$$\nabla \cdot D(\mathbf{r})\nabla\phi(\mathbf{r}) + k^2(\mathbf{r})\phi(\mathbf{r}) = f(\mathbf{r}) \quad \mathbf{r} \in \Omega \tag{3.10}$$

where ∇ is the gradient operator, f is any source located in the medium, $D(\mathbf{r})$ plays the role of a possibly space-varying real-valued diffusion constant and $k^2(\mathbf{r})$ represents the "wave-vector." To accommodate lossy media k^2 will in general be complex. Three classes of boundary conditions are typically considered:

$$\text{Dirichlet} \qquad \phi(\mathbf{r}) = b_D(\mathbf{r}) \quad \mathbf{r} \in \partial\Omega \tag{3.11}$$

$$\text{Neumann} \qquad \frac{\partial\phi(\mathbf{r})}{\partial n} = b_N(\mathbf{r}) \quad \mathbf{r} \in \partial\Omega \tag{3.12}$$

$$\text{Robin (or mixed)} \quad \alpha\phi(\mathbf{r}) + \frac{\partial\phi(\mathbf{r})}{\partial n} = b_R(\mathbf{r}) \quad \mathbf{r} \in \partial\Omega \tag{3.13}$$

where and $\partial\phi/\partial n \equiv \hat{n}\cdot\nabla\phi(\mathbf{r})$ is the projection of the gradient with $\hat{n}$ the outward unit normal of the region. For open problems in three dimensions, the field $\phi(\mathbf{r})$ is required to satisfy the *Sommerfeld radiation boundary conditions* [30, page 87]

$$r\phi(\mathbf{r}) \text{ bounded and } r\left[\frac{\partial\phi(\mathbf{r})}{\partial r} - ik\phi(\mathbf{r})\right] \to 0 \text{ as } r \to \infty \tag{3.14}$$

where $r = |\mathbf{r}|$.

It is useful to think of (3.10) plus the boundary condition as a space domain equivalent to (3.1). Given an input function f and the relevant initial or boundary values, the left hand sides of both (3.1) and (3.10) represent the differential equations that the output must satisfy. Moreover, the ability to solve for $g(t)$ in (3.1) though the use of convolution with an impulse response, has a direct analog in the case of (3.10) where the kernel of the integral is commonly referred to as a Green's function. Unlike the time domain problem where the role of the initial conditions (always zero) is somewhat innocuous, the impact of the boundary conditions along with the specific structure of D and k^2 are quite central to finding a Green's function. This makes the analytical as well as numerical determination of the Green's function a difficult problem in general and one for which a detailed treatment is not really required here.

To demonstrate the applicability of this model, in Table 3.1, we summarize the particular structure it assumes for a variety of application areas. The information in this table is by no means exhaustive. For example, the acoustics entry implicitly assumes a loss-free medium. In the case of electromagnetics, we have not taken into consideration vector effects or the possibility that the magnetic permeability might be space varying. Such generalizations are certainly possible, but will change the precise mathematical form of the quantities D and k^2. Additional details can be found in the references. The physical quantities represented in the Table 3.1 for each application are as follows:

3.3.2 Green's Functions

In the context of inverse problems, it is typically the case that Green's functions are derived for problems with significant symmetry. The quantity D is generally constant and k^2 is a function of at most one spatial variable, say z in Cartesian coordinates or radius, r, in cylindrical or spherical coordinates, thereby resulting in a layered medium. In these cases, the Green's function is usually obtained for open domain problems. Certainly the most straightforward Green's function to derive

Application	ϕ	$\mathbf{D}$	$\mathbf{k^2}$	Reference
Acoustics	Pressure	$1/\rho(\mathbf{r})$	$\omega^2/c(\mathbf{r})^2$	[18, § 1.2]
Scalar Electromagnetics	Electric field	1	$\omega^2\mu_0\left(\epsilon(\mathbf{r}) + i\frac{\sigma(\mathbf{r})}{\omega}\right)$	[50, § 1.5]
Electrostatics	Electric Potential	$\sigma(\mathbf{r})$	0	[14]
Photothermal NDE	Thermal Wavefield	$\kappa(\mathbf{r})$	$i\omega\rho(\mathbf{r})c(\mathbf{r})$	[64]
Diffuse Optical Tomography	Diffuse photon density wavefield	$1/3\mu_s'(\mathbf{r})$	$\mu_a(\mathbf{r}) - i\frac{\omega}{c}$	[1]

Acoustics	ρ = density c = speed of sound
Scalar Electromagnetics	μ_0 = magnetic permeability of free space ϵ = electrical permittivity σ = electrical conductivity
Electrostatics	σ = electrical conductivity
Photothermal NDE	κ = thermal conductivity ρ = density c = specific heat
Diffuse optical tomography	μ_s' = reduced optical scattering coefficient μ_a = optical absorption coefficient c = speed of light

Table 3.1: Form of Helmholtz equation for specific applications. In all cases ω represents angular frequency of the assume time-harmonic excitation.

and the one use predominately in practice is for the case where both D and k^2 are constant. For simplicity, let us assume that $D = 1$. In this case, the Green's function is defined as the solution to the Helmholtz equation satisfying the Sommerfeld condition for a point source located at $\mathbf{r} = \mathbf{r}'$:

$$\nabla^2 g(\mathbf{r}, \mathbf{r}') + k^2 g(\mathbf{r}, \mathbf{r}') = -\delta(\mathbf{r} - \mathbf{r}') \tag{3.15}$$

where the minus sign on the source is a convention followed in *e.g.* the electromagnetics community and the dependence of g on the source location as well as r is always made explicit.

To solve for $g(\mathbf{r}, \mathbf{r}')$ *in 3D and for this problem specifically* it is useful to take advantage of two important facts. First, there is no loss in generality in assuming that $\mathbf{r}' = 0$. Intuitively, because k^2 does not depend on space the solution will only depend on the location of $\mathbf{r}$ relative to $\mathbf{r}'$, *i.e.* $\mathbf{r} - \mathbf{r}'$. Hence we may as well take $\mathbf{r}' = 0$, derive the solution as a function of $\mathbf{r}$ and then replace $\mathbf{r} \leftarrow \mathbf{r} - \mathbf{r}'$. Second, because the source $\delta(\mathbf{r})$ is isotropic (the same no matter which direction we look), $g(\mathbf{r})$ will also be isotropic. That is, when viewed in spherical coordinates, it will depend only on the radial variable, $r = |\mathbf{r}|$, and not the two angular ones. Using these facts and writing the radial part of the Laplacian in spherical coordinates implies that g must be the solution to

$$\frac{1}{r}\frac{\partial^2}{\partial r^2}\left[rg(r)\right] + k^2 g(r) = \delta(r). \tag{3.16}$$

For $r \neq 0$ we can multiply (3.16) through by r and conclude that the function $rg(r)$ must satisfy

$$\frac{d^2}{dr^2}\left[rg(r)\right] + k^2\left[rg(r)\right] = 0 \Rightarrow rg(r) = Ce^{ikr} \Rightarrow g(r) = C\frac{e^{ikr}}{r}.$$

To find the constant C, we integrate the Helmholtz equation over a sphere of radius ϵ around the origin and take the limit $\epsilon \to 0$:

$$\int_{r<\epsilon} \left[\nabla^2 g(r) + k^2 g(r)\right] d\mathbf{r} = -\int \delta(r) d\mathbf{r}.$$

Now, the right hand side is minus one times the integral of a delta function over all space, or just -1. To find C we note first that

$$k^2 \int g(r) d\mathbf{r} = 4\pi k^2 \int_0^\epsilon r^2 g(r) dr \tag{3.17}$$

which goes to 0 as $\epsilon \to 0$. Next, by Gauss' theorem,

$$\int d\mathbf{r} \nabla^2 g = \oint dS \hat{r} \cdot \nabla g = 4\pi r^2 \frac{dg}{dr}\Big|_{r=\epsilon} \tag{3.18}$$

where the middle integral is over the surface of the sphere. Substituting (3.16) for g and simplifying yields $-4\pi C = -1 \Rightarrow C = \frac{1}{4\pi}$. Hence in terms of the case where $\mathbf{r}' \neq 0$ we the constant medium Green's function is

$$g(\mathbf{r}, \mathbf{r}') = \frac{1}{4\pi}\frac{e^{ik|\mathbf{r}-\mathbf{r}'|}}{|\mathbf{r} - \mathbf{r}'|}. \tag{3.19}$$

Moreover because (3.10) is a linear differential equation, by superposition, we know that the field for an arbitrary source $f(\mathbf{r})$ is

$$\phi(\mathbf{r}) = (Gf)(\mathbf{r}) = \int g(\mathbf{r}, \mathbf{r}')f(\mathbf{r}')\mathbf{r}' = \int g(\mathbf{r} - \mathbf{r}')f(\mathbf{r}')d\mathbf{r}'. \tag{3.20}$$

That is, a three dimensional convolution. In other words just as the time invariance of the coefficients in (3.1) led to a convolution, the independence of D and k^2 on space here lead to a convolution-type of input-output relationship.

To make this point clearer and also provide some insight as to how Green's functions can be computed for more general problems, consider the geometry for which the fields are required for all $\mathbf{r} = (x, y, z)$ only where $z > 0$ and at $z = 0$ we require the field to be zero. This structure is one example of a halfspace geometry where symmetry remains in the variables x and y, but is broken for z. In electromagnetic applications the zero boundary condition on the field (an example of a Dirichlet condition, (3.11)) may be thought of as having a perfectly electrically conducting lower halfspace. For acoustics, the requirement of zero field corresponds to a sound-soft lower halfspace. The use of Neumann or Robin conditions also lead to tractable results for the Green's function and are considered in the problems.[1] Finally there are a number of methods for calculating Green's functions where the fields are required in both the top and the bottom halfspaces; however such techniques go well beyond the scope of the material of interest here.

To obtain the Green's function for the Dirichlet condition at the interface we use the method of images. Recall that the Green's function is desired for all $\mathbf{r}$ and $\mathbf{r}'$ in the upper halfspace, $z > 0$. To qualify as a Green's function for this problem all we need a function that satisfies the Helmholtz equation for $z > 0$ and is equal to zero at $z = 0$. Consider the function g_{dhs} (for Dirichlet half space)

$$g_{dhs}(\mathbf{r}, \mathbf{r}') = g(x, y, z, x', y', z') - g(x, y, z, x', y', -z'). \tag{3.21}$$

This function corresponds to placing two sources into a homogeneous medium; one at x', y', z' and a mirror source at $x', y', -z'$. By linearity, g_{dhs} satisfies the Helmholtz equation. By inspection, it is equal to zero at $z = 0$. Thus it must be the Green's function for the problem. [2]

It should be noted that this halfspace problem is still symmetric in x and y because the only variation in the medium are in the z direction. Hence it is easily verified that g_{dhs} is a function only of $x - x'$ and $y - y'$. That is, the homogeneity of the problem in these two dimensions yields a Green's function with a convolutional form in these two variable. For z, the first term in g_{dhs} is of the form $z - z'$, but the second term is a function of $z + z'$ and hence possesses a structure slightly more general than just a convolution in z.

3.3.3 The Inverse Problems

Associated with the Helmholtz equation (as well as its cousins in acoustics, electromagnetics, transport and so forth) are two broad classes of inverse problems: inverse source and inverse scattering. For the inverse source problem we passively acquire field data at the boundary of a region of space and attempt to recover internal structure of a source of field located within the medium. Many

[1] EXERCISE: Neumann halfspace Green's function
[2] EXERCISE: Green for reflecting boundary

astronomy imaging problems have this flavor; for example recovering the structure of a point source such as a star based on data acquired after the fields have passed though the Earth's turbid atmosphere. Similarly, passive sonar problems associated with in ocean acoustic propagation can be cast in this framework. [3]

In contrast to inverse source problems, for inverse scattering problems, we are able to actively excite the medium with the hope of developing an estimate of the space (and perhaps time) varying properties of the medium. The use of seismic sources and hydrophone detectors to image the Earth's sound speed as a function of space is but one example of an inverse scattering problem encountered in geophysical exploration [8]. Inverse scattering problems are encountered in fields such as nondestructive evaluation where low frequency electromagnetic waves are used to excite eddy currents in a material sample from which we seek a map of the electrical conductivity [11]. In medical imaging, near infrared radiation is employed to excite diffuse photon density waves in tissue so that we can image the spatial structure of the tissue's optical absorption and scattering parameters [74]. From this information, tissue oxygenation can then be inferred in order to localize cancerous tumors or detect brain activity.

Mathematically, the inverse source and inverse scattering problems are quite distinct. Referring to (3.10)), the inverse source problem basically requires the determination of $f(\mathbf{r})$ from observations of ϕ assuming the medium parameters, D and k^2 are known. Conversely, the inverse scattering problem allows us to manipulate the sources with the objective of recovering the unknowns D and/or k^2. As we now discuss, the inverse source problem is linear while the inverse scattering problem is nonlinear in the relationship between the data and the desired quantities.

The linearity of the inverse source problem for $f(\mathbf{r})$ follow directly from equation (3.20. While the right hand side of this equation with its convolution form is true only for the problem where the medium is spatially homogeneous, the middle part of the equation is true for any Green's function. That is the fields observed are the integral of the sources against the Green's function. For example, g_{dhs} (3.21) could be used in place of $g(\mathbf{r}, \mathbf{r}')$ if we were dealing with a halfspace problem with a zero Dirichlet condition at $z = 0$. When the medium *is* homogeneous, (3.20) is a multidimensional deconvolution problem whose structure is quite similar to that of (3.2). From this mathematical similarity however one should not conclude that the difficulty of these problems is comparable. For (3.2), one generally has access to the time series $g(t)$ for all time. On the other hand, for most inverse source problems, we are provided data on a restricted subset of the boundary of the medium. Almost never (except in the case of image restoration) do we have a "full data" situation.

Heuristically at least we can count dimensions to see the difference in these problems. For the temporal problem we have full time domain data in the form of $g(t)$ (one data dimension) to determine the structure of a temporal signal, $f(t)$ (one object dimension). The parity here implies very roughly that there is at least some hope that the deconvolution problem solvable. For the inverse source problem in the best case the observed data is collected on a full surface (two dimensions) while we seek the spatial structure of f in three space dimensions. Thus, there is one more degree of freedom in the object space than is provided by the input. Thus intuitively, one would expect that in general, for the inverse source problem, the data will not provide sufficient information to determine the entire structure of the source function. In general, this is true. More technically, in the inverse source literature, one find frequent study of reference made to so-called

[3] More examples would be nice

invisible sources or *non-radiating sources*. These are non-zero functions $f(\mathbf{r})$ which give rise to zero fields on the boundary of the medium. That is, a non-radiating source is a function that lives in the nullspace of the operator $(Gf)(\mathbf{r})$. That this operator typically has a nullspace in turn is a consequence of the dimensional deficit in the data space. Finally, while temporal problems may well have nullspace issues, the availability of full data allow for the use of straightforward Fourier analysis methods from which the nullspace of the operator is readily understood in terms of the spectrum of the impulse response of the system. Indeed, the zeros $H(\omega)$ define the nullspace.

As discussed previously, the goal of the inverse scattering problem is to determine the spatial (and perhaps temporal) structure of $D(\mathbf{r})$ and $k^2(\mathbf{r})$ given boundary field data collected in response to a number of applied sources. To bring out the most salient issues associated with this problem, here we restrict discussion to a special (although still widely applicable) problem: estimation of k^2 assuming it is static as a function of time. Development of an analytical model begins with the decomposition of $k^2(\mathbf{r})$ into two components a background part, $k_b^2(\mathbf{r})$ and a perturbation (or scatter component) $k_s^2(\mathbf{r})$:

$$k^2(\mathbf{r}) = k_b^2(\mathbf{r}) + k_s^2(\mathbf{r}). \tag{3.22}$$

The background component is intended to represent the nominal structure of k^2 which is assumed known *a priori* to the inverse process. Thus the inverse problem amounts to finding the perturbation from the background, $k_s^2(\mathbf{r})$. In practice, $k_b^2(\mathbf{r})$ is chosen such that the Green's function associated with the background-only problem is easily computable. For example, taking k_b^2 to be constant yields a homogeneous medium for which the Green's function is given by (3.19). To see why this is useful, we rewrite the Helmholtz equation as

$$\left[\nabla^2 + k_b^2(\mathbf{r})\right]\phi(\mathbf{r}) = f(\mathbf{r}) - k_s^2(\mathbf{r})\phi(\mathbf{r}) \tag{3.23}$$

plus the associated boundary conditions.

Say we know the Green's function $g(\mathbf{r}, \mathbf{r}')$ for the background problem *i.e.* the problem where $k_s(\mathbf{r}) = 0$. In this case we can integrate both side of (3.23) against the Green's function to obtain

$$\phi(\mathbf{r}) = \int g(\mathbf{r}, \mathbf{r}')f(\mathbf{r}')d\mathbf{r}' - \int g(\mathbf{r}, \mathbf{r}')k_s^2(\mathbf{r}')\phi(\mathbf{r}')d\mathbf{r}'. \tag{3.24}$$

We make the observation that the first term on the right hand side of (3.24) is just the field that is generated when $k_s^2 = 0$. By definition though, this must be the background field (also known as the incident field) which we label $\phi_b(\mathbf{r})$. The second term on the right hand side of (3.24) is the scattered field that arises because of the presence of the perturbation (or scatterer) in the medium. This field we denote by $\phi_s(\mathbf{r})$. Note that in the scattered field integral $k_s^2(\mathbf{r})\phi_s(\mathbf{r})$ plays the same role as $f(\mathbf{r})$ in the background field calculation. Thus we interpret $k_s^2(\mathbf{r})\phi(\mathbf{r})$ as a secondary source which exists in the medium.

The physical intuition behind the mathematics is that field in the medium, $\phi(\mathbf{r})$ on the left hand side of (3.24) is due to two components: a background due to the applied source $f(\mathbf{r})$ and a scattered term due to the perturbation in the medium. The strength of this secondary source (how "bright" it is) is proportional to the amount of field over the source and the contrast of the perturbation. Clearly though, we have a problem here since $\phi(\mathbf{r})$ exists on the right and left sides

of (3.24). That is, ϕ is defined implicitly by this model. To determine the field explicitly, we move the scattered field component to the left hand side of (3.24) which we now write as

$$\left[I + \mathcal{G}_{k_s^2}\right] \phi(\mathbf{r}) = \phi_b(\mathbf{r}) \tag{3.25}$$

where

$$\mathcal{G}_{k_s^2}(\phi)(\mathbf{r}) = \int g(\mathbf{r}, \mathbf{r}') k_s^2(\mathbf{r}') \phi(\mathbf{r}') d\mathbf{r}' \tag{3.26}$$

Eq. (3.25) is known as the Lippman-Schwinger integral equation. Thus we see that the total field ϕ is formally given by applying the inverse of $I - \mathcal{G}_{k_s^2}$ to both side of (3.25)

$$\phi(\mathbf{r}) = \left[I + \mathcal{G}_{k_s^2}\right]^{-1} (\phi_b)(\mathbf{r}). \tag{3.27}$$

To complete the derivation of the inverse scattering problem for $k_s^2(\mathbf{r})$ we substitute (3.27) into (3.24) and rearrange to get

$$\phi(\mathbf{r}) - \phi_b(\mathbf{r}) = \phi_s(\mathbf{r}) = \int g(\mathbf{r}, \mathbf{r}') k_s^2(\mathbf{r}') \left\{ \left[I + \mathcal{G}_{k_s^2}\right]^{-1} (\phi_b) \right\} (\mathbf{r}') d\mathbf{r}'. \tag{3.28}$$

Thus, assuming that the incident field is known perfectly, we can subtract it from the total field to arrive at the "data" for this problem, observations of scattered field. These observations are related to the desired unknown $k_s^2(\mathbf{r})$ trough the scattered field integral on the right side of (3.28). It is important to note however that this relationship is quite nonlinear. While the explicit presence of k_s^2 in the integral *is* linear, the operator $I + \mathcal{G}_{k_s^2}$ also depends on k_s^2. Hence so too will its inverse thereby resulting in an overall model that is nonlinear in k_s^2.

This type of dependence significantly complicates the inverse problem. For linear problem such as deconvolution, X-ray tomography, and the inverse sorce problem, the full power of the vector space ideas can be brought to bear both on the analysis of and solution to the problem. Moreover, as discussed in the next chapter, these tools also yield a rich variety of explicit analytical inversion formulae for recovering the unknown given the data. When the problem is not linear, the mathematical tools required to obtain both useful and deep insight as well as fast inversion methods are less developed. Those that do exist (such as the $\bar{\partial}$ method of [82] are based on mathematical principles whose level of sophistication and abstraction is well beyond that of linear vector spaces.

Much more common for the nonlinear problems however is the use of the physical model in the context of an optimization problem. The estimated k_s^2 is that function which extremizes (maximizes or minimizes depending on the context) a cost function which includes a term requiring fidelity to the data. Thus the tools of mathematical optimization and numerical analysis (for solving the forward problem) play a key role in the development of these methods. Chapter 6 will provide details on this topic.

Perhaps the most common approach to dealing with inverse scattering problems is linearization. If $k_s^2(\mathbf{r})$ is small in amplitude relative to the background and nonzero over a region of space small relative to the wavelength of the fields probing the medium, then the first Born approximation is known to be valid. Physically, this form of interaction of fields with perturbations is known as *diffraction* and the problem of recovering k_s^2 from such data is termed *diffraction tomography*.

While there are many ways of deriving the Born approximation [18, 47], the opertator theoretic approach to the inverse scattering problem we have pursued here suggests a very straightforward

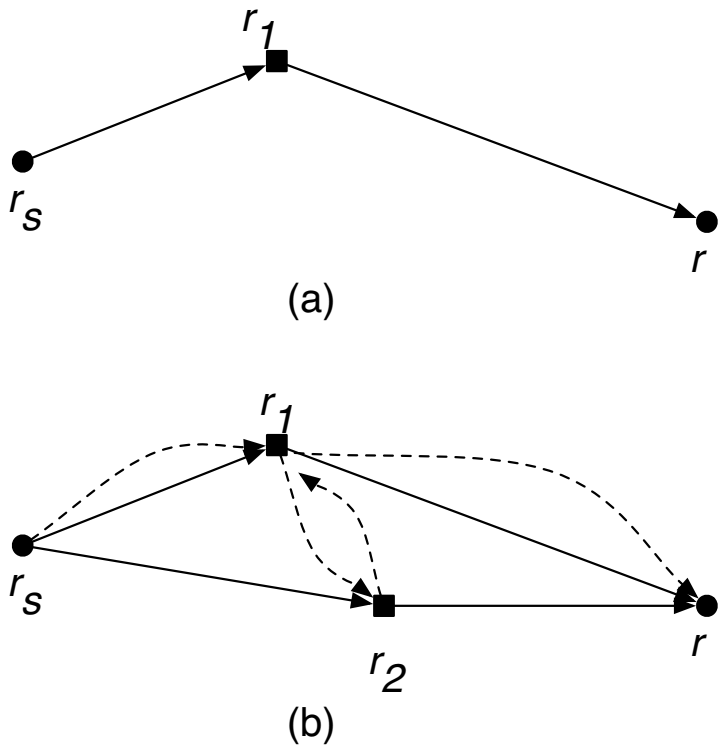


Figure 3.3: Depiction of (a) single and (b) multiple scatter interactions of a field with a pertubation

technique. The Lippman-Schwinger equation requires the inverse of the operator $I-\mathcal{G}_{k_s^2}$. Formally, we can use the scalar geometric series $(1+a)^{-1} = 1-a+a^2-a^3\ldots$ valid for $|a|<1$ in the current setting to arrive at:

$$\left[I-\mathcal{G}_{k_s^2}\right]^{-1} = I-\mathcal{G}_{k_s^2}+\mathcal{G}_{k_s^2}\mathcal{G}_{k_s^2}-\ldots. \tag{3.29}$$

Retaining only the first term in (3.28) and noting that $(I\phi_b)(\mathbf{r}) = \phi_b(\mathbf{r})$ results in the first Born model

$$\phi_s(\mathbf{r}) \approx \int g(\mathbf{r},\mathbf{r}')k_s^2(\mathbf{r}')\phi_b(\mathbf{r}')d\mathbf{r}' \tag{3.30}$$

which *is* linear in the unknown $k_s^2(\mathbf{r})$. One can derive so-called n-th order Born models by retaining n terms in the series. It is easy enough to verify that the resulting model has a dependence on k_s^2 which is polynomial in nature with terms up to and including $\left[k_s^2(\mathbf{r})\right]^n$.

Physically, the model in (3.30) captured the first order interaction of the fields with the pertubation in k^2 and hence is often referred to as a single scatter approximation. To understand this concept a bit more intuitively, consider the case where $k_s^2(\mathbf{r}) = a\delta(\mathbf{r}-\mathbf{r}_1)$, a single point like reflector with "reflectivity" a and the source is a point source at $\mathbf{r}_s$. In this case the incident field is $G(\mathbf{r},\mathbf{r}_s)$ and the scattered field is then $G(\mathbf{r},\mathbf{r}_1)aG(\mathbf{r}_1,\mathbf{r}_s)$. This is expression can be physically interpreted as follows. First, the source creates the field $G(\mathbf{r},\mathbf{r}_s)$. Second this field interacts with

the only scatterer in the medium, the object at $\mathbf{r}_1$. Roughly speaking, interaction here means that the scatterer responds to the incident field by setting up a secondary field. The strength of that field is the product of the amount of field provided by the initial source at the location of the scatterer, $G(\mathbf{r}_1, \mathbf{r}_s)$, and the reflectivity of the scatterer itself, a. Because the scatterer is acting a point source , the spatial distribution of the resulting field must be that of the Green's function for the problem for a source located at $\mathbf{r}_1$. Hence we see the term $G(\mathbf{r}, \mathbf{r}_1)$. Pictorially, this interaction is captured in Fig. 3.3(a). The source creates field. The field scatters from the perturbation. The resulting scattered field is observed at $\mathbf{r}$. We denote this first order interaction as $\mathbf{r}_s \rightarrow \mathbf{r}_1 \rightarrow \mathbf{r}$

Now consider a slightly more complicated example where $k_s^2(\mathbf{r}) = a\delta(\mathbf{r} - \mathbf{r}_1) + b\delta(\mathbf{r} - \mathbf{r}_2)$. The scattered field now is just the superposition of the contributions from each of the two components of k_s^2, $G(\mathbf{r}, \mathbf{r}_1)aG(\mathbf{r}_1\mathbf{r}_s) + G(\mathbf{r}, \mathbf{r}_1)bG(\mathbf{r}_2, \mathbf{r}_s)$. This situation is shown in Fig. 3.3(b) with the solid lines which indicates the observed field as the sum of the contributions from the two scatterers. Not captured by this model though are the multiple-interactions (known as *multiple-scatter*) of the incident fields with objects at $\mathbf{r}_1$ and $\mathbf{r}_2$. For example, the second order interactions $\mathbf{r}_s \rightarrow \mathbf{r}_1 \rightarrow \mathbf{r}_2 \rightarrow \mathbf{r}$ (dashed lines in Fig. 3.3(b)) and $\mathbf{r}_s \rightarrow \mathbf{r}_2 \rightarrow \mathbf{r}_3 \rightarrow \mathbf{r}$ are not taken into account. Higher order cases follow easily. In fact, it is not hard to show that n-th order scattering can be exactly captured by including n terms from (3.29) thereby obtaining the n-th order Born approximation.

3.4 Discretization Methods

Much of this manuscript is concerned with the solution to discretized forward and inverse problems. As seen in this chapter, the continuum forms of the models of interest fall into one of two classes: partial/ordinary differential equations or integral equations. The study of computational methods for these problems represents a field of work in its own regard with a vast and highly interesting literature [43, 51, 79, 80]. Here we provide a brief introduction to a couple of the more common and easily implementable discretization techniques and provide citations to references containing more thorough treatments of each approach.

The material in this section is devoted primarily to discretization methods for the Helmholtz models in § 3.3. In the case of the convolutional methods in § 3.1, signal processing oriented texts such as [77] provide highly readable and quite useful coverage of a variety of sampling methods. The ordinary differential equations literature e.g. [27] contains extensive coverage of advanced techniques, such as Runge-Kutta, for solving ODEs. Finally, one method for discretizing the Radon transform model, (3.9), is discussed in the problems at the end of this chapter. [4] More generally, the methods of moments, which we cover as a tool for discretizing the Helmholtz equation, can be used here as well. See [43] for an example.

Finite Difference Methods

Perhaps the most straightforward means of producing a discrete representation of a partial differential equation is through the use of finite differences. Assuming the functions we wish to approximate are sufficiently smooth, the basic idea behind finite difference is to approximate derivatives using a

[4] EXERCISE: Build Radon model w/ partial voluming

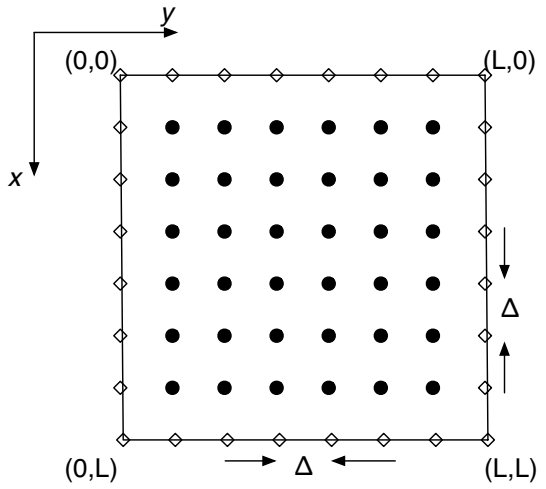


Figure 3.4: Two Dimensional Square Grid

Taylor series expansion of the function up to second order

$$f(x_0+\delta) = f(x_0) + \delta\frac{df(x_0)}{dx} + \frac{\delta^2}{2}\frac{d^2f(x_0)}{dx^2} + O(\delta^3) \tag{3.31}$$

where *e.g.* $df(x_0)/dx$ is the first derivative of f evaluated at x_0 and $O(\delta^3)$ mean that the error in the value of $f(x_0+\delta)$ goes top zero as the cube of δ.

For the problems of interest here, we require derivatives up to second order. By evaluating (3.31) at $+\delta$ and $-\delta$ and performing a bit of algebra yields the following central difference approximation[5] to the first derivative:

$$\frac{df(x_0)}{dx} = \frac{f(x_0+\delta) - f(x_0-\delta)}{2\delta} + O(\delta^3). \tag{3.32}$$

where the notation $O(\delta^n)$ indicates that the remaining terms in the power series are all of the form $a_k\delta^k$ for $k \geq n$. Similarly, to obtain an $O(\delta^2)$ approximation to the second derivative of f we can use

$$\frac{d^2f(x_0)}{dx^2} = \frac{f(x_0+\delta) - 2f(x_0) + f(x_0-\delta)}{\delta^2} + O(\delta^2). \tag{3.33}$$

Equations (3.32) and (3.33) provide really all that is necessary to obtain a discrete representation for the Helmholtz equation and its associated boundary conditions.

As a specific example, consider first the case where we want to solve

$$[\nabla^2 + k^2]\phi = 0 \tag{3.34}$$

on the two dimensional $L \times L$ square shown in Fig. 3.4 subject to the boundary conditions

$$-\frac{\partial\phi(x,0)}{\partial y} = b_{N,left}(x) \qquad \frac{\partial\phi(x,L)}{\partial y} = b_{N,right}(x) \tag{3.35}$$

$$\phi(0,y) = b_{D,top}(y) \qquad \phi(L,y) = b_{D,bot}(y) \tag{3.36}$$

[5]May want to add material on forward and backward difference as well

To keep the development simple, we break the square into an $N_x \times N_y$ grid of points spaced uniformly in x and y with grid spacing Δ. In Figure 3.4, $N_x = N_y = 8$. We shall refer to these points using one of two ordering schemes. On the one hand, it may be convenient to label quantities location at the point i rows from the top and j columns from the left side of the grid as *e.g.* $\phi(x_i, y_j)$ or just $\phi_{i,j}$ for $i = 1, 2, \ldots, N_x$ and $j = 1, 2 \ldots, N_y$. Alternatively, we may lexicographically order the points by "stacking" one row on top of the other and refer to $\phi(\mathbf{r}_n)$[6] or ϕ_n where $n = 1, 2, \ldots, N_x N_y$ and is related to i and j via $n = i + (j-1)N_x$. [7]

Recalling that in 2D,

$$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}$$

we can use (3.33) at any point (i, j) in the grid to discretize (3.34) as

$$\frac{\phi_{i-1,j} - 2\phi_{i,j} + \phi_{i+1,j}}{\Delta^2} + \frac{\phi_{i,j-1} - 2\phi_{i,j} + \phi_{i,j+1}}{\Delta^2} + k_{i,j}^2 \phi_{i,j} = 0. \tag{3.37}$$

In theory, this formula provides $N_x \times N_y$ linear equations in $N_x \times N_y$ unknowns. Switching to n ordering we can gather the $\phi_{i,j}$ into a vector

$$\mathbf{x} = \begin{bmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_{N_x N_y} \end{bmatrix}.$$

Eq. (3.37) then can be used to define a matrix $\mathbf{A}$ and a vector $\mathbf{b} = \mathbf{0}$ such that $\mathbf{A}\mathbf{x} = \mathbf{b}$. The entries of $\mathbf{A}$ along any given row will consist of 1's, -2's, and k_n^2 [8] in order to affect the summation with the corresponding elements of $\mathbf{x}$ implied by (3.37).

This simple idea fine for all of the points (i, j) corresponding to solid circles in Fig. 3.4; however for the open circles, (3.37) requires knowledge of points outside of the grid. For points on the left edge of the grid $i = 1, 2, \ldots, N_x$ and $j = 1$ we need values for $\phi_{i,0}$. On the right edge we require ϕ_{i,N_y+1}. On the top and bottom, $\phi_{0,j}$ and $\phi_{N_x+1,1}$ are called needed. To accommodate these points we make use of the boundary conditions.

For the top and bottom of the grid, the Dirichlet conditions indicate exactly what the field values should be. This information can be used in the equations governing the fields values one row from each of these sides. In the case of the second row of points, $i = 2$, for $j = 2, 3, \ldots, N_y - 1$ (3.37) is

$$\frac{\phi_{1,j} - 2\phi_{2,j} + \phi_{3,j}}{\Delta^2} + \frac{\phi_{2,j-1} - 2\phi_{2,j} + \phi_{2,j+1}}{\Delta^2}. + k_{2,j}^2 \phi_{2,j} = 0.$$

but the boundary condition is $\phi_{1,j} = b_{D,top}(y_j)$. Hence after some algebra we get

$$\phi_{3,j} + \phi_{2,j-1} + \phi_{2,j+1} + (-4 + \Delta^2 k_{2,j}^2)\phi_{2,j} = -b_{D,top}(y_j). \tag{3.38}$$

[6] With a slight abuse of notation, here we let $\mathbf{r}$ represent a point in two dimensions whereas previously we have been using it to denote a point in 3D.

[7] EXERCISE: $n \to i, j$ and 3D

[8] EXERCISE: Show this

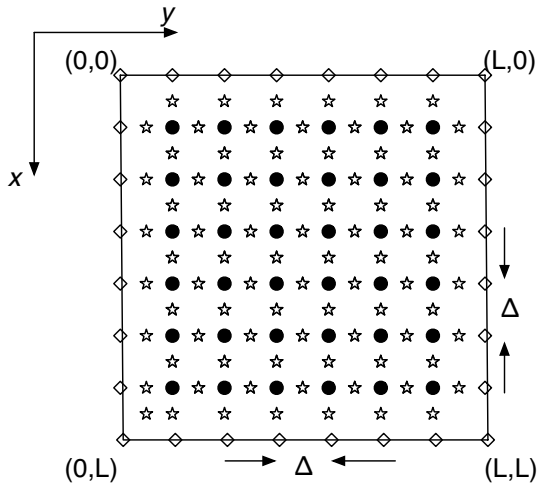


Figure 3.5: Two Dimensional Square Grid for Inhomogeneous Diffusion

Eq. (3.38) and its counterpart for all points along row $N_x - 1$ are then used in place of (3.37) for building the appropriate rows of the $\mathbf{A}$ matrix and $\mathbf{b}$ vector.

For the left and the right sides of the grid we use the centered approximation of the first derivative

$$-\frac{\partial\phi(0,y)}{\partial y} \approx \frac{\phi_{i,0} - \phi_{i,2}}{2\Delta} \quad \text{and} \quad \frac{\partial\phi(L,y)}{\partial y} \approx \frac{\phi_{N_x+1,j} - \phi_{N_x-1,j}}{2\Delta}$$

and the Neumann boundary conditions to conclude

$$\phi_{i,0} = \phi_{i,2} + 2\Delta \times b_{N,left}(x_i) \tag{3.39}$$

$$\phi_{i,N_y+1} = \phi_{i,N_y-1} + 2\Delta \times b_{N,right}(x_j) \tag{3.40}$$

As with the discussion surrounding (3.38), eqs. (3.39) and (3.40) can be used to eliminate $\phi_{0,j}$ and $\phi_{N_x+1,j}$ from (3.37) for points $(i,j) = (1,j)$ and $(i,j) = (N_x,j)$ resulting in alternations to the corresponding rows of $\mathbf{A}$ and elements of $\mathbf{b}$ [9].

Within the context of the Helmholtz model, one further complications to the use of finite differences comes though the introduction of an inhomogeneous diffusion coefficient $D(\mathbf{r})$ in (3.10). Expanding the first term on the left hand side of (3.10) in Cartesian coordinates we obtain:

$$\nabla \cdot [D(\mathbf{r})\nabla\phi(\mathbf{r})] = \frac{\partial}{\partial x}D(x,y,z)\frac{\partial}{\partial x}\phi(x,y,z) + \frac{\partial}{\partial y}D(x,y,z)\frac{\partial}{\partial y}\phi(x,y,z) + \frac{\partial}{\partial z}D(x,y,z)\frac{\partial}{\partial z}\phi(x,y,z) \tag{3.41}$$

One way of achieving a finite difference discretization of this differential operator is to assume that the samples of $D(\mathbf{r})$ exist on a grid that is offset by $\Delta/2$ from that used for ϕ and k^2 [1]. This is

[9] EXERCISE: Build $\mathbf{A}$ & $\mathbf{b}$

shown in Fig. 3.5 where we reproduce the grid used previously this time using the stars to identify the points where $D(\mathbf{r})$ is evaluated. The use of first order differences yields for the x term in (3.41)

$$\frac{\partial}{\partial x}D(x,y,z)\frac{\partial}{\partial x}\phi(x,y,z) \approx \frac{1}{\Delta}\left\{D(i+\frac{1}{2},j)\left[\frac{\phi(i+1,j)-\phi(i,j)}{\Delta}\right]-D(i-\frac{1}{2},j)\left[\frac{\phi(i,j)-\phi(i-1,j)}{\Delta}\right]\right\} \quad (3.42)$$

with similar formulae holding for the y and z components.

In our discussion of finite difference methods, we have been very careful to consider only closed-domain problems where the Dirichlet, Neumann, and Robin conditions are imposed on the boundaries. Infinite domain problems can certainly be addressed though the use of finite difference methods. However because the domains for these problems are infinite and computers have finite memory it is necessary to terminate the grid in some manner that makes the resulting finite dimensional system reproduce the true physics. This is not at all easy. For example, one might be tempted to solve the problem on a large grid and just set the fields at the boundary (now well removed from where we are really interested in observing the fields) to zero. This strategy produces difficulties for two reasons. First, we must solve a problem larger than necessary which means there is a large computational cost to this naive approach. Second, and more troubling, is that the use of this imposed boundary condition does not replicate the physics. Rather than solving an open problem, we are solving a problem in a large "box" with imposed Dirichlet conditions. The two are not the same and as a result the accuracy of the solution will suffer, perhaps considerably.

Thus, there has been extensive work in the development of *absorbing boundary conditionss* (ABCs). ABCs are methods for terminating the grid in ways that eliminate or at least minimize the numerical artifact. Some of these methods are based on the Sommerfeld radiation condition [73], but many other approaches have also been put forth [85]. Ultimately, building robust ABCs that work for wide classes of possible problems is far from easy and well beyond the scope of this manuscript.

A last element of the finite difference method to address is the choice of Δ, the sampling rate. As is well known in the signal processing literature, the Shannon sampling theorem states that one must sample a band-limited temporal signal at a rate at least twice as fast as the highest frequency one expects to be present. Failure to sample at this rate or higher results in artifacts known as aliasing [77, Section 4.2]. The same reasoning can be applied to our problem as well with the result being that one needs a sampling rate, Δ, that is at least half the shortest wavelength (equivalent to twice the highest spatial frequency) expected to be encountered in the solution to the problem. For most applications, the wavelength information is related to the structure of k^2. As a simple example, for the scalar electromagnetic problem detailed in Table 3.1, in the case where $\sigma = 0$ nd ϵ is constant, it is know that the wavelength is

$$\lambda = \frac{2\pi}{\omega\sqrt{\mu_0\epsilon}}.$$

Similar formulae hold for the other problems detailed in Table 3.1.

When the media are inhomogeneous, the determination of the wavelength is not as easy as the use of a formula. In practice this is addressed in two steps. First, one generally has some idea

as to the nominal structure of the region being investigated as was the case when we assumed a background value for $k^2(\mathbf{r})$. This information can be used to place a conservative bound on the minimum wavelength. Second, in the sampling rate used is typically far higher than that which the Shannon theory would prescribe. Practical computational methods for problems like the ones of interest here sample at between 10 and 20 times per wavelength.

As we have seen both through the development here and via the exercises at the end of this chapter, finite difference methods yield a matrix-vector system of linear equations that need to be solved in order to determine the values of the field at the sample points in the grid. These linear systems are noteworthy for two reasons. First, the presence of the sample values of D and k^2 within the entries of the matrix further demonstrates the nonlinearity of the inverse scattering problem. Indeed, if we let $\mathbf{k^2}$ and $\mathbf{d}$ represent the vectors of lexicographically ordered samples of $k^2(\mathbf{r})$ and $D(\mathbf{r})$ then the finite difference system may be written

$$\mathbf{A}(\mathbf{k^2}, \mathbf{d})\mathbf{x} = \mathbf{b} \Rightarrow \mathbf{x} = \mathbf{A}^{-1}(\mathbf{k^2}, \mathbf{d})\mathbf{b}. \tag{3.43}$$

The nonlinear dependence of $\mathbf{A}^{-1}$ on the elements of $\mathbf{k^2}$ and $\mathbf{d}$ means that the data we have for the inverse problems, the samples of the field located $\mathbf{x}$ also are related in a nonlinear manner to these desired unknowns.

The second feature of $\mathbf{A}$ is its sparse structure. Because $\mathbf{A}$ arises from the discretization of a differential operator, very few elements on any given row differ from 0. Thus, even while $\mathbf{A}$ may have a huge number of rows and columns, the number of nonzero elements that need to be stored in a computer are much more modest. To exploit this structure in solving for $\mathbf{x}$, one commonly does not build $\mathbf{A}^{-1}$ explicitly. Rather, there exist a collection of techniques for finding $\mathbf{x}$ that generate a sequence of iterates $\mathbf{x}^{(0)}, \mathbf{x}^{(1)}, \ldots$ converging to $\mathbf{A}^{-1}\mathbf{b}$ and requiring only routines that compute the product of $\mathbf{A}$ and $\mathbf{A}^T$ with input vectors. The sparse structure of $\mathbf{A}$ implies that such operations can be done quite efficiently. These solution techniques are known as Krylov subspace methods and essentially represent a class of algorithms that generalize the well known conjugate gradient method to cases where $\mathbf{A}$ is not symmetric and positive definite. Details on their implementation and use can be found in [79].

Method of Moments

While finite differences are most easily employed in the solution of closed domain partial differential equations, open domain problems where the Green's function is known or can be calculated are well suited to discretization by the method of moments. We concentrate here on the application of this approach to the Lippman-Schwinger integral equation (3.25) and the Born approximation (3.30).

The method of moments is based on the representartion of a function as a linear combination of a set of pre-defined basis elements:

$$f(\mathbf{r}) = \sum_{i=1}^{N} f_i b_i(\mathbf{r}) \tag{3.44}$$

where the f_i are unknown coefficients and the b_i are the basis functions. A variety of common

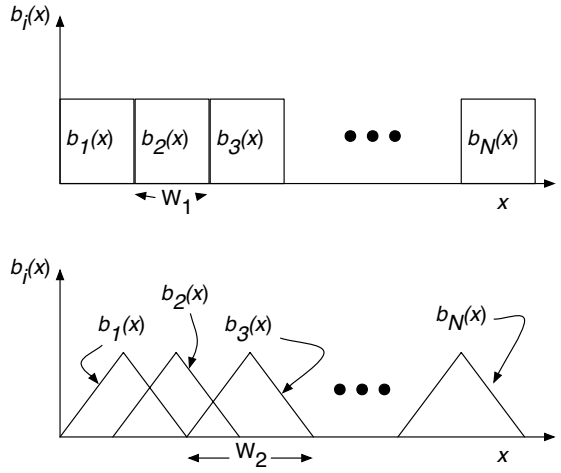


Figure 3.6: Piecewise constant and piecewise linear basis functions for method of moments

choices are made for the $b_i(\mathbf{r})$. In some cases, one may choose to use Fourier-type of functions:

$$b_i(x,y,z) = \sin(mx)\sin(ny)\sin(pz)$$
$$b_i(x,y,z) = \cos(mx)\cos(ny)\cos(pz)$$

where $m = 1, 2, \ldots N_m$, $n = 1, 2, \ldots N_n$, and $p = 1, 2, \ldots N_z$ and there is some ordering scheme like the one discussed previously for mapping a triple (m, n, p) into an index i. Illustrated in Fig. 3.6 are a couple of popular choices for one dimensional basis functions. In the top of the figure are functions capable of producing piecewise constant (stair-step) representations of $f(\mathbf{r})$. A piecewise linear possibility is shown in the bottom of Fig. 3.6.

Multi-dimensional basis functions can be obtained as separable combinations of the one dimensional elements; *i.e.* $b_i(\mathbf{r}) = b_i(x,y,z) = b_n(x)b_m(y)b_p(z)$ where again we assume an ordering capable of taking the triple (m, n, p) to the single index i. It is easy enough to verify that in two dimensions, the separable basis obtained by using the piecewise constant basis at the top of Fig. 3.6 corresponds to typical image pixels that have a value of one on a rectangular region and are zero everywhere else. [10]

Within the context of inverse methods, these basis functions are used for two discretization tasks. On the one hand they are frequently employed in nonlinear inverse methods as a means of numerically solving the Lippman-Schwinger equation, (3.24) for the fields ϕ. Additionally, basis function expansion is a common tool in the reduction of general linear inverse problems to forms amenable to solution on a computer.

To discretize the Lippman-Schwinger equation in a manner well suited to the ultimate problem

[10] EXERCISE: Plot basis elements in 2D and 3D

of determining k_s^2 we start by assuming basis expansions for $\phi(\mathbf{r})$, $\phi_b(\mathbf{r})$, and $k_s^2(\mathbf{r})$

$$\phi(\mathbf{r}) = \sum_{i=1}^{N_1} \phi_i b_i(\mathbf{r}) \tag{3.45}$$

$$\phi_b(\mathbf{r}) = \sum_{j=1}^{N_2} \phi_{b,j} c_j(\mathbf{r}) \tag{3.46}$$

$$k_s^2(\mathbf{r}) = \sum_{k=1}^{N_3} f_k d_k(\mathbf{r}). \tag{3.47}$$

Using (3.45) – (3.47) in side of (3.24) yields

$$\sum_{i=1}^{N_1} \phi_i b_i(\mathbf{r}) = \sum_{j=1}^{N_2} \phi_{b,j} c_j(\mathbf{r}) - \sum_{i=1}^{N_1} \phi_i \tilde{g}_i(\mathbf{r}) \tag{3.48}$$

where

$$\tilde{g}_i(\mathbf{r}) = \sum_{k=1}^{N_3} f_j \int g(\mathbf{r}, \mathbf{r}') b_i(\mathbf{r}') d_k(\mathbf{r}')\, d\mathbf{r}'.$$

Reduction of (3.47) to a matrix vector problem is achieved by taking the inner product of both sides of (3.48) with $b_i(\mathbf{r})$. Specifically, we have in matrix form

$$\mathbf{B}\boldsymbol{\phi} = \mathbf{C}\boldsymbol{\phi}_b - \mathbf{G}\boldsymbol{\phi} \tag{3.49}$$

where $\boldsymbol{\phi}$ is the length N_1 column vector of coefficients in the expansion of $\phi(\mathbf{r})$ in (3.45), similarly for $\boldsymbol{\phi}_b$ and the elements of $\mathbf{B}$, $\mathbf{C}$ and $\mathbf{G}$ are

$$\mathbf{B}_{i,j} = \int b_i(\mathbf{r}) b_j(\mathbf{r})\, d\mathbf{r} \quad i = 1, 2, \ldots, N_1, j = 1, 2, \ldots N_1$$

$$\mathbf{C}_{i,j} = \int b_i(\mathbf{r}) c_j(\mathbf{r})\, d\mathbf{r} \quad i = 1, 2, \ldots, N_1, j = 1, 2, \ldots N_2$$

$$\mathbf{G}_{i,j} = \int b_i(\mathbf{r}) \tilde{g}_j(\mathbf{r}) \quad i = 1, 2, \ldots, N_1, j = 1, 2, \ldots N_1.$$

Finally, assuming $\mathbf{B}$ is invertible (which follows if and only if the $\phi_i(\mathbf{r})$ are linearly independent)[11], we can solve for the unknown vector $\boldsymbol{\phi}$ as[12]

$$\boldsymbol{\phi} = (\mathbf{B} + \mathbf{G})^{-1} \mathbf{C}\boldsymbol{\phi}_b \tag{3.50}$$

To use these basis functions in the discretization of linear inverse problems (deconvolution, X-ray tomography, inverse source and Born-based inverse scattering) requires two pieces of mathematical

[11] EXERCISE: Prove

[12] EXERCISES: Implement for pixel basis

preparation. First, we begin by noting that all linear inverse problems of interest here take the form of an integral equation relating the data, $g(\mathbf{r})$ to an object $f(\mathbf{r})$

$$g(\mathbf{r}) = \int K(\mathbf{r}, \mathbf{r}')f(\mathbf{r}')d\mathbf{r}' \tag{3.51}$$

where we have been a bit cavalier in our denotation of the independent variables of the object, $\mathbf{r}'$, and those of the data, $\mathbf{r}$. For deconvolution, these quantities are both scalar. For X-ray tomography, the input variables are space, $\mathbf{r}' = (x, y)$ while the output variables are those of the Radon transform $\mathbf{r} = (t, \theta)$. In the case of inverse source and linearized inverse scattering, $\mathbf{r}$ ranges over the region of space where the data are collected while $\mathbf{r}'$ takes values in the volume (for 3D) or image plane (in 2D) where f is defined.

Second, (3.51) is written in terms of the continuum variables $\mathbf{r}$. In reality, the data available for inversion are sampled functionals of these quantities. For simplicity, we assume here these functionals are linear[13]. To be consistent with the mathematical structure developed in Chapter 2 and which we shall use later in this manuscript, this implies the existence of a vector space, X, holding our objects, $f(\mathbf{r})$ as well functionals $l_i\, i = 1, 2, \ldots, N$ mapping this space into components of a data vector. For now, we shall take X to be the Hilbert space of square integrable functions defined over an appropriate domain, $-\infty$ to $+\infty$ for temporal problems and some compact set of $\mathbb{R}^2$ or $\mathbb{R}^2$ for spatial inverse problems. The simplest such functional is that which "samples" $f(\mathbf{r})$ at $\mathbf{r}_i$, the location of the i-th sensor

$$g_i = (l_i|g) = \int \delta(\mathbf{r}_i - \mathbf{r}')g(\mathbf{r}')d\mathbf{r}'. \tag{3.52}$$

More generally, this linear functional approach provides sufficient flexibility to handle sensors which perform weighted integrals of the continuum data over finite sized apertures:

$$g_i = (l_i|g) = \int w_i(\mathbf{r})g(\mathbf{r})d\mathbf{r} \tag{3.53}$$

where the weighting functions are (a) sufficiently well behaved to assure the existence of the integral (b) typically functions of $\mathbf{r} - \mathbf{r}'$, and (c) typically sharply peaked around the location of the i-th sensor, $\mathbf{r} = \mathbf{r}_i$.[14]. Making use of the general model (3.51), results in

$$\begin{aligned} g_i &= \int \left(K(\mathbf{r}, \mathbf{r}')|l_i(\mathbf{r})\right) f(\mathbf{r}')d\mathbf{r}' \\ &= \int d\mathbf{r}' \left[\int d\mathbf{r}\, w_i(\mathbf{r})K(\mathbf{r}, \mathbf{r}')\right] f(\mathbf{r}'). \end{aligned} \tag{3.54}$$

[13]Nonlinear functionals do play an important role in many applications. For example, in many optical imaging problems the sensors do not provide the full complex electric field $\phi(\mathbf{r})$, but only its magnitude integrated over a "pixel" in a CCD array [35, page 49]. That is the i-th data point is well approximated as

$$g_i = \int_{i-\text{th pixel}} |\phi(\mathbf{r})|^2 d\mathbf{r}.$$

[14]EXERCISE: Gaussian and diff. of Gaussian

Upon substitution of (3.44) into (3.54) and rearranging we obtain

$$g_i = \sum_j K_{i,j} f_i \tag{3.55}$$

$$K_{i,j} = \int w_i(\mathbf{r}) K(\mathbf{r}, \mathbf{r}') b_j(\mathbf{r}') \, d\mathbf{r} \, d\mathbf{r}'. \tag{3.56}$$

Finally, arranging the g_i and f_j into vectors and the $K_{i,j}$ into matrices yields the final form of the discretized linear model

$$\mathbf{g} = \mathbf{K}\mathbf{f} \tag{3.57}$$

There are a two issues that arise in consideration of the use of the method of moments. First, creating the matrices $\mathbf{B}$, $\mathbf{C}$, $\mathbf{G}$ and $\mathbf{K}$ requires the evaluation of multi-dimensional integrals. With judicious choices for the various basis functions, some of these calculations can be done in closed form (see exercises). In many cases, analytical expressions for these integrals do not exists so numerical procedures must be employed. There are a variety of methods for performing these calculations and a large number of canned software packages available for use. We refer the reader to [21] and [76, Chapter 4] for more information. Second, the various matrices and vectors associated with both the discretized Lippman-Schwinger and linear forward model are often complex. As we show in the exercises at the end of this chapter, by separating the real and imaginary components, it is possible to still obtain a matrix-type of relationship between the coefficients of the input object and the output samples. [15]

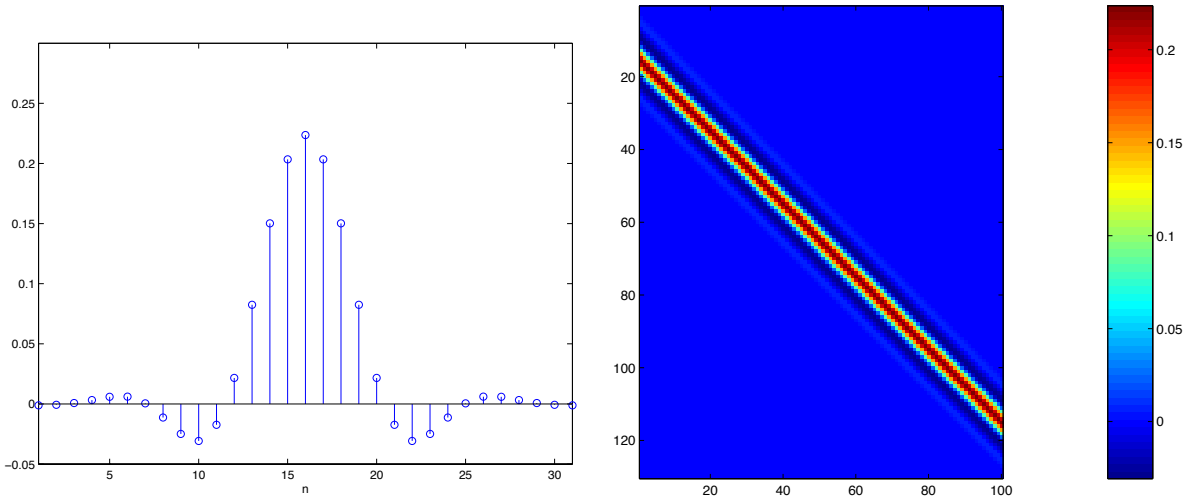
Toeplitz Matrices and Deconvolution

Before leaving the issues of discretization, we touch briefly on the issue of matrices associated with shift invariant, that is, convolutional, problems. There are two reasons for this discussion. First, the matrices associated with these problems have a clearly identifiable structure which is easily captured both mathematically as well as visually. Thus, these matrices provide a nice (perhaps quintessential) example by which we can begin to build up some intuition about inverse problems by actually looking at their components. Second, as indicated above, these matrices have signficant structure. While it is somewhat beyond the scope of this manuscript, a key issue associated with numerically solving inverse problems arising in two and three dimensional applications is computational complexity. Writing down a solution is one thing. Implementing it in some reasonably efficient manner is quite another. As discussed in greater detail in [48, 88], the structure inherent in these matrices leads to extremely efficient algorithmic implementations.

In one dimension, a convolutional operator is one for which the kernel $K(t, s)$ is a function not of t and s separately, but of their difference, $t - s$. Most methods for discretizing such integral equations result in matrices where this structure is preserved in that the element on row m and column n is not a function of m and n individually, but of $m - n$. So $\mathbf{K}_{m,n} = \mathbf{K}_{m-n}$. As an example, consider a problem

$$g(t) = \int K(t - s) f(s) \, ds$$

[15] EXERCISE: Re & Im system



(a) Samples of impulse response for Toeplitz matrix example

(b) A Toeplitz matrix arising from the discretization of a convolution problem

Figure 3.7: 1D discrete convolution matrix

which, upon discretization yields

$$g(n) = \sum_m K(n-m)f(m).$$

Let us suppose that the discrete impulse response contains 31 coefficients, plotted in Fig. 3.7(aa), while the sampled form of $f(t)$ contains 100. The goal now is to build a matrix $\mathbf{K}$ such that $\mathbf{Kf} = \mathbf{g}$ where $\mathbf{f}$ and gmb are vectors containing $g(n)$ and $f(n)$. For a discrete convolution problem with an impulse response of length M and input signal of length N, the length of the output vector will be $M + N - 1$ [77, Section 5.3.1]. Hence $\mathbf{g} \in \mathbb{R}^{119}$ so $\mathbf{K}$ will be of size 130×100. Due to the convolutional structure of the problem, $\mathbf{K}_{m,n} = \mathbf{K}_{m-n} = K(m-n)$ where we assume $K(n)$ is equal to zero for $n < 0$ and $n > 31$.

In Figure 3.7(b), we display an "image" of the resulting discretized convolutional operator. That is, the pixel on row m and column n of the image is color-coded according to the value of $\mathbf{K}_{m,n}$. Under this scheme an identity matrix would appear as an image with a single narrow stripe running from the top left corner to the bottom right corner. Visually, the $m - n$ dependence is seen in the "stripes" along the diagonals. Matrices possessing this structure are called *Toeplitz* [88, Chapter 5]. Moreover, we see by looking at this matrix that its action on a vector will be local in the sense that the nonzero structure on any given row of $\mathbf{K}$ is restricted to a fairly small portion of the full 100 possible elements.

In two dimensions, Toeplitz-like structure is also seen, but the situation is a bit more complex. A discrete two dimensional convolution takes the form

$$g(i,j) = \sum_{p,q} K(i-p, j-q)f(p,q). \tag{3.58}$$

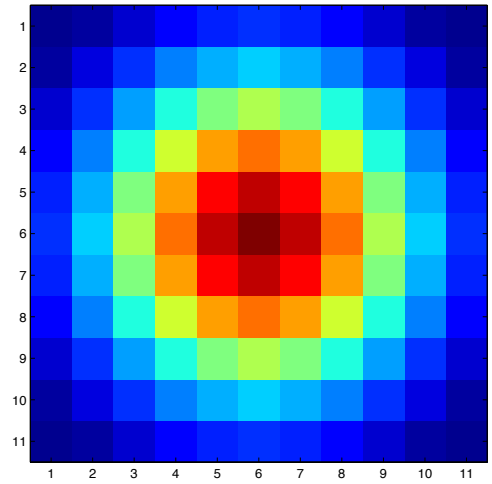


Figure 3.8: Kernel for 2D Convolution Problem

We would like to write (3.58) in the standard matrix-vector form $\mathbf{g} = \mathbf{K}\mathbf{f}$. To do so requires that we establish an ordering which maps row-column indices such as (i, j) or (p, q) to a single index n or m for the column vectors $\mathbf{g}$ and $\mathbf{f}$. Here we employ what is known as a lexicographical ordering scheme whereby the vector $\mathbf{f}$ is obtained by "stacking" the columns of $f(p, q)$ one on top of the other. Assuming that $p = 1, 2, \ldots, R$ and $q = 1, 2, \ldots C$ then the following transformations between (p, q) and $m = 1, 2, \ldots RC$ ordering are readily verified for this ordering method:

$$(p, q) \text{ to } m: \qquad m = R(q-1) + p \tag{3.59}$$

$$m \text{ to } (p, q) \qquad q = 1 + \left\lfloor \frac{n-1}{R} \right\rfloor \tag{3.60}$$

$$p = n - R(q-1) \tag{3.61}$$

For this ordering scheme, the matrix $\mathbf{K}$ takes on form known as *block Toeplitz with Toeplitz blocks* (BTTB) [88, Chapter 5]. This term is best understood via an example. A two dimensional 11×11 blurring kernel is shown in Fig. 3.8. Generalizing the 1D case, assuming an 15×15 input image, the resulting output image will be of size $11 + 15 - 1 \times 11 + 15 - 1$ or 25×25. The resulting matrix $\mathbf{K}$ will be of size $25^2 \times 15^2 = 625 \times 225$ and is shown in Fig. 3.9. The lexicographic ordering of the input and output images result in a block structure for $\mathbf{K}$. Each block is of size 25×15 and is itself a Toeplitz matrix. Additionally, the blocks themselves have a Toeplitz pattern. Thus, the $(1, 1)$ block of 25×15 elements is the same as blocks $(2, 2)$, $(3, 3)$ etc. Similarly blocks $(1, 2)$, $(2, 3)$, $(3, 4)$, are equal.

3.5 Exercises

3.1 Often times in programming solutions to multidimensional inverse problems, it is necessary to generate a unique ordering of the pixels in an image or voxels in a discretized volume. Say we have an image represented by an array of numbers, `I(m,n)` where $1 \leq m \leq M$ and

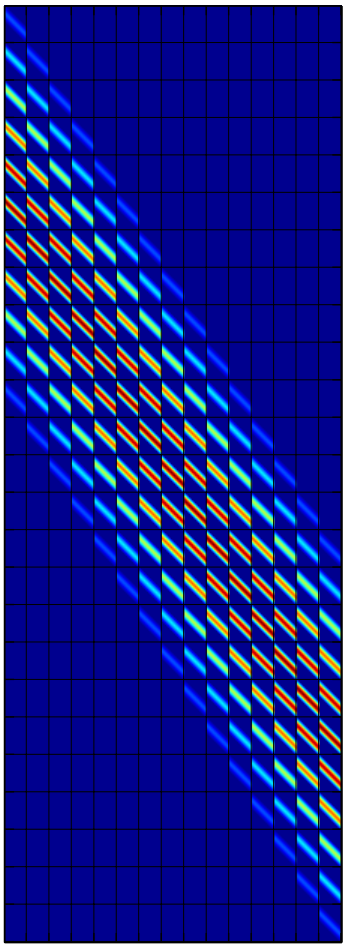


Figure 3.9:

$1 \leq n \leq N$. In Matlab, we can "transform" this image into a vector via the colon operation `I1D = I(:)` and access elements of `I1D` using the index `i` which runs from 1 to $N \times M$. Note that ":" basically stacks the columns of `I`. See Figure 3.10 for an example of how this works in the case of a 4×4 image.

(a) To access the pixels in an image one-by-one we can either write a pair of nested `for` statements that loop over columns and rows (or rows and columns) or we can write a single `for` loop that iterates over the index i. Which ordering, rows-then-columns or columns-then-rows gives provides the access to the elements of the image as is induced by the Matlab colon operator?

(b) Write a pair of functions which translate between `(m,n)` and `i`. The first function should take as input only `m`, `n`,`M`, and `N` and return the corresponding value of `i` which is consistent with the ordering imposed by ":." The second function should take only `i`,`M`,

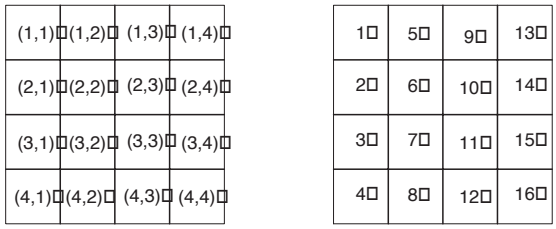


Figure 3.10: Example of index transformation in 2D. Row-column pixel indices are sorted into a linear array of numbers from 1 to 16.

and `N` and return the corresponding `m` and `n`. For full credit, these functions should be as simple as possible. Verify that these functions work.

(c) In Matlab, the command `A = rand(3,2,4)` creates a 3D "volume" of size $3 \times 2 \times 4$ filled with random numbers. As with images, `A1D = A(:)` is a stacked form of the values in `A`. Now, voxels indexing requires three quantities: `m`, `n`, and `p`. Repeat the first two parts of this problem now for the 3D case. Note that doubly nested `for` loops become triply nested loops etc.

3.2 In this problem we consider a model arising in the use of thermal waves for non-destructive evaluation. In 1D, the thermal wavefield, $u(x)$, for a time-harmonic source can be shown to satisfy the ordinary differential equation in the absence of sources

$$\left(\frac{d}{dx^2} - \sigma^2\right) u(x) = 0$$

where $\sigma = \sqrt{i\omega/\alpha}$, ω is the frequency of operation, and α is known as the thermal diffusivity.

(a) Define the Green's function, $G(x, x_0)$ for this problem as the solution to the ODE

$$\left(\frac{d}{dx^2} - \sigma^2\right) G(x, x_0) = -\frac{1}{\alpha}\delta(x - x_0).$$

Argue that the most general form of $G(x, x_0)$ is $Ae^{\sigma x} + Be^{-\sigma x}$. Where A and B are constants whose values are to be determined from the boundary conditions.

1. It can be show that G is continuous across x_0 as a function of x, but has a discontinuous first derivative. Show that for a homogeneous medium in which G must go to zero as x goes to $\pm\infty$, we have

$$G(x, x_0) = \frac{1}{2\alpha\sigma} e^{-\sigma|x - x_0|}$$

(b) Now determine G for the case of a semi-infinite medium where $G(0, x_0) = 0$ and as $x \to \infty$, $G(x, x_0) \to 0$.

3.3 An alternate way of linearizing the Helmholtz equation is via the Rytov approximation. While the end results is quite similar to Born, the small differences make Rytov far more

widely applicable than Born (see the next problem set). Supressing explicit dependence on r, consider the PDE for the field u

$$(\nabla^2 + k^2)u = 0.$$

Supose we assume that the solution to this equation takes the form

$$u(r) = e^{\phi(r)}.$$

where ϕ is a generally complex valued phase function.

(a) Show that ϕ satisfies

$$(\nabla\phi)^2 + \nabla^2\phi + k_0^2 = -o(r)$$

where $o(r) = k_0^2(r) - k_s^2(r)$ is the object function we wish to image.

Similar with the Born approximation, we decompose the phase function, ϕ into a background and scattered part: $\phi = \phi_b + \phi_s$ where the background phase function is by definition the phase function associated with the background field, $u_b(r) = e^{\phi_b(r)}$.

(b) Since the background field solves the background scattering problem, argue that

$$k_b^2 + (\nabla\phi)^2 + \nabla^2\phi = 0$$

(c) Show now that ϕ_s satisfies

$$2\nabla\phi_0 \cdot \nabla\phi_s + \nabla^2\phi_s = -(\nabla\phi_s)^2 - o(r)$$

(d) Assuming that $u_0(r)$ is a plane wave so that $\nabla^2 u_0 = k_0^2 u_0$ and using the fact that $\nabla^2(u_0\phi_s) = \nabla((\nabla u_0)\phi_s + u_0(\nabla\phi_s))$ show that

$$(\nabla^2 + k_0^2)(u_0\phi_s) = -u_0\left[(\nabla\phi_s)_o^2(r)\right]$$

(e) Using the above result, conclude that under a certain assumption we have$\phi_s(r) = \frac{u_B(r)}{u_0(r)}$ where $u_B(r)$ is the Born scattering function

$$\int G(r, r')u_0(r')o(r')dr'$$

What is the required assumption? This is the Rytov aproximation.

1. In practice, we do no measure the Rytov phase, but rather the total field, $u(r)$. Assume we can perfectly subtract the incident field, u_0, to effectively measure the scattered field, $u_B(r)$. Show that the Rytov forward model is, in terms of quantities we measure or know,

$$u_0(r)\ln\left(\frac{u_s(r)}{u_0(r)} + 1\right) = \int G(r, r')u_0(r')o(r')dr'$$

3.4 Here we want to develop a Born-like approximation for the problem

$$\nabla \cdot \sigma(r)\nabla\phi(r) + k^2(r)\phi(r) = 0 \tag{3.62}$$

where both σ and k^2 can vary.

1. Prove that
$$\int_V a(r)\left[\nabla \cdot (b(r)\nabla c(r))\right] dr = -\int_V \left[\nabla a(r) \cdot \nabla c(r)\right] b(r) dr$$
where r is a point in three-space and V is the volume of integration, a, b and c are all differentiable, and $b = 0$ on the boundary of V.
2. Let us assume in (2.2) that $\sigma(r) = \sigma_0 + \sigma_s(r)$ and $k^2(r) = k_0^2 + k_s^2(r)$ where σ_0 and k_0^2 (the background material parameters) are both constants. Show that for k_s and σ_s "small" we have to first order that the scattered field can be written
$$\phi_s(r) = -\int_V \nabla_{r'} G(r, r') \cdot \nabla\phi_0(r')\sigma_s(r')dr' - \int_V G(r, r')\phi_0(r')k_s^2(r')dr' \tag{3.63}$$
where G is the Green's function for the background problem.
3. Explain how the above result can be used in the contaxt of an inverse problem.

3.5 Here we want to build a routine for constructing the system matrix for a Born-based inverse problem at least for a simplified problem. The basic setup for the problem is shown in Figure 3.11. We assume that k_s^2 is restricted to be nonzero over a very thin region of space of size $\delta \times Y \times Z$ meters with δ very, very small and centered at the point x_0.

The Green's function for this problem is

$$G(r, r') = \frac{e^{ik_0|r-r'|}}{4\pi|r - r'|}$$

where $|r - r'| = \sqrt{(x - x')^2 + (y - y')^2 + (z - z')^2}$. For this because δ is so small you may assume that $x = x' = x_0$. Also, k_0 can be an arbitrary complex number.

Point source transmitters (blue dots) and point receivers (orange dots) are arrayed at arbitrary locations around the square d meters from the edge. The data for the inverse problem will be comprised of observations of the scattered field for all source-receiver pairings. More formally, the ith source is taken to be $a_i\delta(r - r_{S,i})$ where a_i is the amplitude (a real number) and $r_{S,i}$ is the location in 3D of this point source. Each source gives rise to a scattered field, $\phi_s(r)$ which we observe at position $r_{R,j}$. Here $i = 1, 2, \ldots, N_S$ and $j = 1, 2, \ldots, N_R$. The data obtained for this problem is the $N_S \times N_R$ vector of complex valued scattered fields associated with all source-receiver combinations.

1. Write a first kind linear integral equation relating $k_s^2(r)$ to the scattered field data:
$$\phi_s(r_j) = \int_V K(r_j, r')k_s^2(r')dr' \tag{3.64}$$

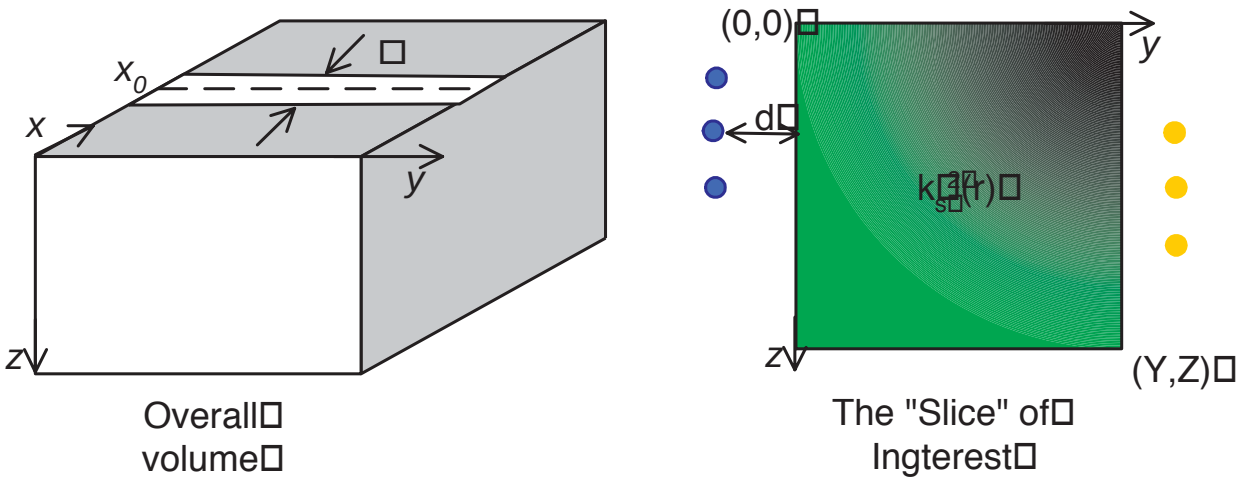


Figure 3.11: Setup for inverse scattering type of problem

2. Say we break space up into an $N_Z \times N_Y$ array of pixels. Develop a method for discretizing (2.4) by sampling r' at the center of each pixel.
3. Write a piece of code which takes as input the following
   - x_0 and δ in meters
   - Z and Y in meters
   - Integers N_z and N_Y
   - The N_S 3D locations of all sources (positions all in meters)
   - The N_R 3D locations of all receivers (positions all in meters)
   - The complex number k_0 in meters^{-1}.

   and produces as output the $N_S N_R \times N_Z N_Y$ system matrix that maps the pixel values for k_s^2 into the data vector.
4. One row of your matrix is a function of two variables since $r' = (y', z')$ and thus may be thought of as an $N_Z \times N_Y$ "image." Explain why we can think of each pixel in that image as the sensitivity of the datum for that row to a unit change in the corresponding pixel of k_s^2.
5. Generate the matrix for the case where
   - $x_0 = 1$ and $\delta = .001$
   - $Y = 1$ and $Z = 1$
   - $N_z = N_y = 40$
   - N_S=1 and the location is $(x_S, y_S, z_S) = (1, -0.1, 0.5)$
   - The $N_R = 1$ and the location is $(x_R, y_R, z_R) = (1, 1.1, 0.5)$

- The complex number $k_0^2 = (1)^2$.

Your K matrix now has only a single row. Plot an "image" of the real and imaginary parts of this row. Explain any structure you see.

6. Repeat the above but for k_0 taking on the following values: $\{10, 25, 10 + .1 * \sqrt{-1}, 10 + \sqrt{-1}10 + 10 * \sqrt{-1}\}$. Usding the Matlab `imagesc` command, plot the real and imaginary parts of the "image" of the kernel for all of these cases being sure that they all have the same colorscale (see the `caxis` command). How do these changes in the backgorun d wavenumber impact the structure of the matrix? Qualitatively at least, will the resulting inverse problem be better posed or less well posed?

7. Now let us change things a bit. As shown in Figure 3.12, lets us say we put N transducers equally spaced on each side of the 2D region we are imaging. Each transducer acts as a source which produces fields that are measured at all $4N$ point. Thus a single scattering experiment yields a total of $4N \times 4N$ complex valued points of data. Hence assuming we discretize the medium into an array of $N_y \times N_z$ pixels, we could write the model in the form $y_{cx} = K_{cx} f$ where y_{cx} is the length $16N^2$ complex valued data vector and K_{cx} the discretized Born kernel. What is typically done however is to break $K_c x$ and $y_c x$ into real and imaginary components so that the overall model is of the form:

$$y = \begin{bmatrix} \text{real}\{y_{cx}\} \\ \text{imag}\{y_{cx}\} \end{bmatrix} = \begin{bmatrix} \text{real}\{K_{cx}\} \\ \text{imag}\{K_{cx}\} \end{bmatrix} f \equiv Kf$$

Thus the "data" vector is of length $32N^2$ while the system matrix has $32N^2$ rows and $N_y N_z$ columns.

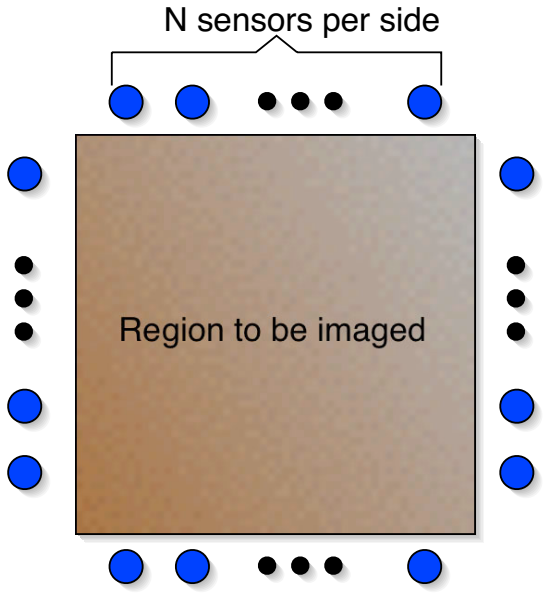


Figure 3.12: Setup for Part (g)

For this part of the problem, please examine the nature of the singular structure of the system matrix as we vary N and k_0. Specifically, plot on a log scale the singular values for the system matrices obtained for all combinations of the following parameter sets:

- $N \in \{1, 5\}$
- The real part of $k_0 \in \{0, 25\}$.
- The imaginary part of $k_0 \in \{0, 10\}$.

For all of these experiments take $N_y = N_z = 30$, $Z = 1$, $Y = 1$, $x_0 = 1$, $d = .1$, $\delta = .01$. Explain how adding sensors and varying the structure of the background medium impact the ability to recover an image.

3.6 A common problem in fields ranging from medical imaging to atmospheric physics is the reconstruction of objects from so-called *projection data*. More precisely, under this schemeAs we have ddiuscussed, for X ray tomography, the data, $g(\theta, t)$, are related to the object, $f(x, y)$, via a linear transformation called the *Radon transform*, which is of the form:

$$g(\theta, t) = (Rf)(\theta, t) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \delta(t - x\cos\theta - y\sin\theta)\, f(x, y)\, dx\, dy \tag{3.65}$$

where $\delta(x)$ is the Dirac delta function.

1. In $x - y$ space, plot $\delta(t - x\cos\theta - y\sin\theta)$. How does this function change as the variable t changes? What about θ? In what sense is $y(\theta, t)$ a *projection* of $f(x, y)$?
2. Derive an explicit expression for $(R^*g)(x, y)$, the adjoint operator for (2.16). What is the physical interpretation of the Radon transform adjoint? Is the Radon transform operator self-adjoint?
3. Based on your previous answers, if we have a *finite-dimensional* representation of the projection operator as the $N_\theta N_t \times N^2$ matrix T, so $y = Tx$, what is the backprojection operation on the data y?
4. For most practical problems, $f(x, y)$ is contained in some finite region of the plane. Assume that this region corresponds to the unit disc centered at the origin. Show that under this assumption the following relationship is true for any reasonably well behaved (specifically, square integrable) function $F(t)$

$$\int_{-1}^{1} g(\theta, t)\, F(t)\, dt = \int_{\text{disc}} f(x, y)\, F(x\cos\theta + y\sin\theta)\, dxdy \tag{3.66}$$

5. Use (2.17) and a judicious choice of $F(t)$ to prove the projection-slice theorem.
6. Now assume that $g(\theta, t)$ is sampled at a finite number of points in θ-t space: $g_{i,j} = g(\theta_i, t_j)$, $i = 1, \ldots, M_\theta$ and $j = 1, \ldots, M_t$. Assume we order these pairs first by t then by θ. Under this discretization scheme nd using the "pixel ordering" methods fromt he first problems set, show that the vector g of projection samples may be related to the vector f of object coefficients f_i via a matrix-vector equation of the form:

$$g = Tf. \tag{3.67}$$

What is the analytical expression for the ijth element of T? Assume that $f(x, y)$ is supported in a square region centered about the origin.

7. On the class web site you will find a tar file **radon.tar.gz** with MATLAB code to perform projection of a field. Given a matrix, **f**, representing a 2-D field and a vector **theta**, holding a collection of angles between 0 and 180 degrees, the function call routine **radon(f,theta)** finds the projection of **f** at the angles specified in **theta**. Note that the number of samples per projection are determined automatically by the routine and is not under the users control. Also in the tar file is a routine **makeb.m**, which can be used to create unit coordinate fields (i.e. fields with zeros everywhere but one pixel location). Using these routines write a MATLAB routine **projmtx.m** that generates the matrix T for a given square array size N and vector of angles. Note: to save space the routines generate output in sparse matrix format, which needs to be converted using **full** for certain MATLAB functions.
8. Using **projmtx.m** generate T for the case of N uniformly spaced angles from 0 to 180 degrees for $N = 4$, 8, 16. Look at the matrices using the function **spy**. Is this a shift-invariant system? What is the dimension of the null space of T for the various N? For N=16, use the MATLAB function **null** to find vectors (images) in the null space of T.

3.7 One common class of inverse problems is concerned with determining the temperature distribution in a region of space at some time in the past based on measurements of the temperature at the present time. Here we will examine the associated forward problem.

The most basic form of the inverse heat conduction problem is to determine the initial temperature on the boundary of a medium (i.e. the temperature at time $t = 0$) from corresponding boundary observations at time $t = T$ with $T > 0$. For simplicity, we will assume that the region of space is a line (i.e. we have a one dimensional problem) extending from $x = 0$ to $x = \pi$ and that we are interested in the dynamics of the problem from time 0 through time T. For our model, the heat propagation is well described using the so-called diffusion equation

$$\frac{\partial u(x,t)}{\partial t} = \frac{\partial^2 u(x,t)}{\partial x^2} \tag{3.68}$$

subject to the boundary conditions

$$u(0,t) = u(\pi,t) = 0 \quad 0 \le t \le T \tag{3.69}$$

and the initial conditions

$$u(x,0) = f(x) \quad 0 \le x \le \pi. \tag{3.70}$$

The function $u(x, t)$ is the temperature on the bar at position x and time t.

(a) Assuming that $u(x, t)$ can be written as $v(x)q(t)$, show that $v(x)$ and $q(t)$ must individually satisfy two constant coefficient ordinary differential equations. What are these equations? Be sure to indicate the initial or boundary conditions for each equation. (Hint: This procedure of solving a partial differential equation is known as *separation of variables* and is thoroughly described in most mathematics texts on partial differential equations or most engineering texts on electromagnetics.)

(b) Your answer from the previous part should indicate that $v(x)$ satisfies an ODE of the form

$$\mathbf{D}v(x) + \lambda v(x) = 0 \tag{3.71}$$

where $\mathbf{D}$ is some differential operator and λ is a constant. The problem is to find *all* $v(x)$ which satisfy this ODE and the associated boundary conditions. For each such solution, there will be a corresponding λ. The solution is known as an eigenfunction and the λ is the associated eigenvalue.

A typical method for solving the ODE for $v(x)$ is to write this function as

$$v(x) = \sum_n a_n \psi_n(x) \tag{3.72}$$

where each of the $\psi_n(x)$ satisfies the ODE as well as the boundary conditions and the a_i are coefficients to be determine later. Using Fourier methods, determine the $\psi_n(x)$ and λ_n for this problem.

(c) Using the results of (b), solve the ODE for $q(t)$. Note that the solution should depend on the index n in (3.72).

(d) The results of (b) and (c) should indicate that $u(x,t)$ is of the form

$$u(x,t) = \sum_n \beta_n q_n(t) \psi_n(x). \tag{3.73}$$

Determine the values of β_n such that (3.73) is satisfied.

(e) The inverse problem of interest is to determine $f(x)$ from observations of $u(x,t)$ taken at time $t = T$. Let $g(x) \equiv u(x,T)$. Show that

$$g(x) = \int_0^\pi K(x,y) f(y) dy. \tag{3.74}$$

by explicitly identifying the SVD of $K(x,y)$.

3.8 Another common inverse problem is that *of numerical differentiation*. The idea here is that the data we are provided, $g(x)$, is related to the object of interest, $f(x)$, via the integral equation

$$g(x) \equiv (\mathbf{A}f)(x) = \int_0^x f(y) dy \tag{3.75}$$

where $x \in [0,1]$ and $y \in [0,1]$.

(a) Identify $K(x,y)$ for this problem. Show that the problem is equivalent to a deconvolution problem. What is the impulse response?

(b) We want to find the SVD for this problem. That is, we want to determine μ_n, $u_n(x)$ and $v_n(y)$ such that

$$K(x,y) = \sum_n \mu_n u_n(x) v_n(y). \tag{3.76}$$

Recall from class that determination of these quantities is closely tied to solving the eigen-problem $\mathbf{A}^*\mathbf{A}u = \mu^2 u$. Thus, we must first determine the operator $\mathbf{A}^*\mathbf{A}$. Show that

$$(\mathbf{A}^* g)(y) = \int_y^1 g(x)dx. \tag{3.77}$$

What is the integral operator for $(\mathbf{A}^*\mathbf{A}f)(x)$?

(c) Using the results of (c), show that any u which satisfies $\mathbf{A}^*\mathbf{A}u = \mu^2 u$ for the numerical differentiation problem also satisfies the ODE

$$\mu^2 \frac{d^2u(x)}{dx^2} + u(x) = 0 \tag{3.78}$$

subject to the boundary conditions $u(1) = u'(0) = 0$ where $u'(x) = du(x)/dx$. Sketch a u which satisfies these boundary conditions.

(d) Solve for all μ_n and $u_n(x)$ which satisfy (3.78).

(e) Using the defining equations for the SVD discussed in class, find $g_n(y)$.

(f) In this and the previous problem, you obtained the SVD for two different integral operators. Each of these operators is a function of two variables and thus may be regarded as an image.

(a) Using MATLAB, provide a mesh plot or image plot for the heat conduction and numerical differentiation kernels. For the heat problem, try $T = 0.1$ and $T = 3$. (Note: you will need to select some value N_{max} at which to terminate the infinite sums. Be sure to indicate how you did this.)

(b) What characteristic of the mesh or image of the differentiation kernel indicates that this is in fact a deconvolution problem?

(c) Sines and cosines are the eigenfunctions for shift invariant (i.e. convolutional) problems. The SVDs for both the heat conduction and differentiation problems were composed of sines and cosines. The plot for the differentiation problem indicated that it was in fact shift invariant. The image for the heat kernel did not share this structure. What is going on?

3.9 Recall the heat conduction problem in which we were interested was the reconstruction of the initial temperature distribution over a region from observations of the temperature at some later time. Letting $f(x)$ be the temperature at time $t = 0$ and $g(x;T)$ be the temperature measured at some time $t = T > 0$, then we showed that the following relationship holds

$$g(x;T) = \int_0^\pi K(x,y;T)f(y)dy \tag{3.79}$$

$$K(x,y;T) = \sum_{n=1}^{\infty} e^{-n^2 T} \sin nx \sin ny \tag{3.80}$$

where $0 \le x, y \le \pi$.

1. Discretize this system using the Galerkin method with flat-top functions for a particular value of T. The infinite sum in (3.80) should be truncated at that N for which the corresponding eigenvalue is $<$ 1e-10. Assuming that we want to sample the highest frequency in the problem using 5 basis functions per wavelength, how does this choice of N affect the number of unknowns in the problem? Show that the end result of the discretization provides a decomposition of the discrete operator as the product of three matrices the middle which is diagonal. Is this an SVD? The result of this problem should be a MATLAB file which produces a matrix representation of K when given an end time, T.
2. Without resorting to the SVD, prove that the discrete problem must have exact nullspace of rank at least 1. (Hints: Look at the operator and consider the boundary conditions.)
3. Now we want to see how this system behaves as we change T for two different sets of initial conditions. Suppose that f_1 is a flat-top function of unit norm which is non-zero over $\pi/3 \leq x \leq 2\pi/3$, f_2 is a flat-top function of unit norm which is non-zero for $\pi/4 \leq x \leq 3\pi/3$ and that we want to look at the system for equally spaced values of time between $T = 0.1$ and $T = 5$.
   (a) To make things easy, determine the finest discretization (i.e. largest values of N and largest number of basis functions) which will be needed for any time, T in the range of interest. For the remainder of the problem, you should build discretized matrices at this resolution.
   (b) For each time instant of interest you should write MATLAB code which build both the K matrix and the data vectors for the two different initial conditions.
   (c) How does the data for these two different sets of initial conditions evolve as time progresses? What about the singular values of the K matrix at each time? Look at the quantity $\delta(T) = \|g_1(x;T) - g_2(x;T)\|$ where $g_i(x;T)$ is the data vector corresponding to f_i. Despite the fact that $\|f_1 - f_2\|$ is constant over time, $\delta(T)$ is changing. What does the behavior of $\delta(T)$ tell you about how difficult it will be to recover f from g as time goes on? How is this supported by the behavior of the singular values?

Chapter 4

Analytic Methods for Linear Inverse Problems

As discussed in the introduction to this manuscript, there are two broad ways of addressing inverse problems. In many circumstances, the physics of the problems is so complex that there is no choice but to start by discretizing the model and using tools of numerical analysis and optimization theory to obtain a solution. On the other hand in a surprisingly broad range of application areas there is sufficient mathematical structure that it is possible to analytically invert the underlying continuum model to obtain either a closed form expression or a step-by-step algorithm for obtaining the continuous object $f(\mathbf{r})$ from generally continuously available data $g(\mathbf{r})$, $g(t,\theta)$, etc. Clearly at the end of the day, implementation of these methods on a computer requires discretization. In those cases where this step can delayed however, significant insight into the problem and deep connections across a wide range of problem areas is achievable.

More specifically, in this chapter we concentrate on a class of problems whose solution can be obtained through the use of Fourier methods. Specifically, in a sense that we make more precise shortly, the Fourier transform of the data can be related in closed form to that of the object. Thus, inversion is quite closely tied to inverting the Fourier transform. Clearly, one dimensional deconvolution is the quintessential example of such a problem. Referring to (3.5), the Fourier transform of the data is the product of the transform of the input and the frequency response of the underlying linear time invariant system. Thus inversion requires only multiplying the data by the reciprocal of the system frequency response and then performing an inverse Fourier transform. In the language of § 3.1 we have

$$f(t) = \frac{1}{\sqrt{2\pi}} \int \frac{G(\omega)}{H(\omega)} e^{i\omega t} d\omega. \tag{4.1}$$

Ignoring for the time being what this means when $H(\omega) = 0$, it turns out that the fundamental structure embodied by (4.1) is at the heart of inverse methods for a wide assortment of spatial inverse problems.

This class of inverse methods share the common structure that the data are weighed integrals of the object observed over what are essentially, linear apertures. This is clearly the case for CAT as discussed in Chapter 3 where the weighting function is a Dirac delta. For applications such

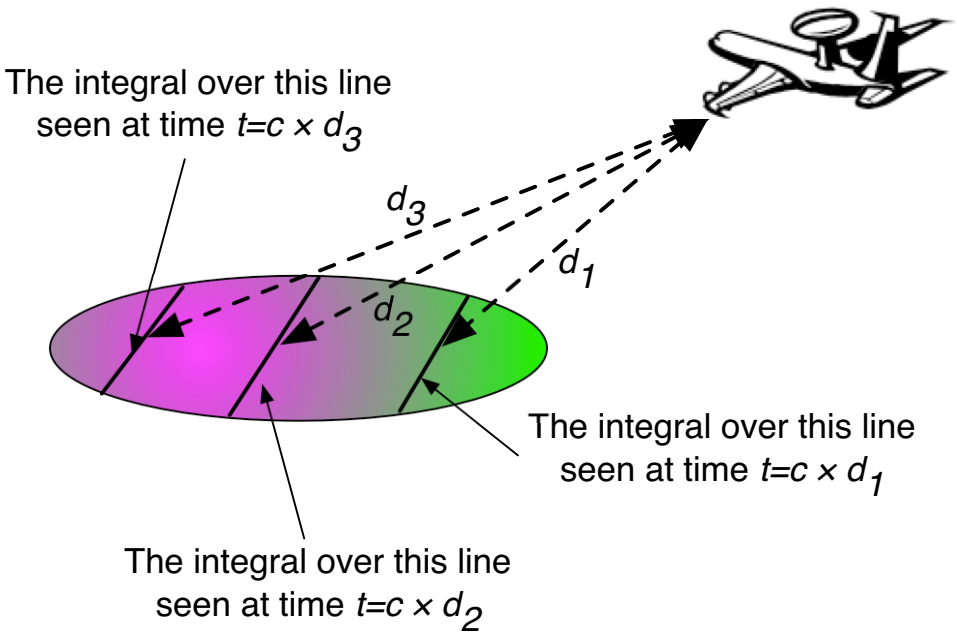


Figure 4.1: Geometry of Synthetic Aperture Radar Problem.

ultrasonic medical imaging and geophysical exploration using electromagnetic and acoustic waves the Born model represents an accurate approximation to the true physics. For these problems, one frequently employs a linear array of receivers to record the fields produced by collection of remote sources [8, 23, 24]. The Born kernel provides the weight here. As a final example, we consider the case of synthetic aperture radar (SAR) imaging, a sensing tool used in geophysical remote sensing and military surveillance [15, 46, 68, 83]. One common approach to SAR has an airborne platform probing with radar pulses as it circles a spot on the earth. As illustrated in Fig. 4.1, the data collected from any one pulse is a time series related to the field backscattered from the earth. Assuming, among other things, a flat earth, each element of that time series is well approximated as an integral of the earth's reflectivity function over a line located $d = tc$ meters from the radar with c the speed of light. Thus, as the aircraft circles, each time series carries essentially the same information as a single CAT projection.

The common underlying Fourier structure that makes CAT, SAR, and Born imaging similar to (4.1) is obtained by analytically relating the one dimensional Fourier transform of each projection to a piece of the two dimensional Fourier transform of the object. As we shall detail shortly, for problem where the data are line integral, such as CAT and SAR, this "piece" is a line though the 2D frequency domain origin. For Born, the line is replaced by a circle. Ultimately then, the collection of projections provides direct observation of a portion of the 2D spatial Fourier transform of f. Thus, inversion can be carried out by inverse Fourier transform, as in (4.1). Alternatively, the structure of these regions can be exploited to obtain another class of algorithms in which each projection is filtered, "smeared" back into the image region, and the smeared results from each

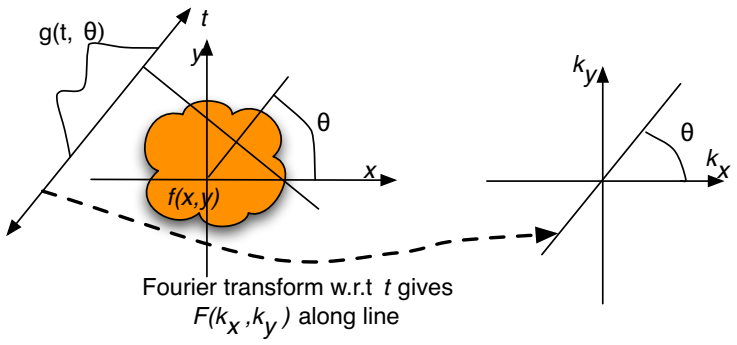


Figure 4.2: The Fourier Slice Theorem for X-Ray Tomography

projection are coherently summed. Within this chapter we discuss both options.

4.1 Inverting the Radon Transform

The first step in developing a closed form inverse formula for the Radon transform is deriving the well known Fourier Slice Theorem. Let us recall that the forward Radon transform defined in (3.9) is

$$g(t,\theta) = \int_{-\infty}^{\infty} f(x,y)\delta(t - x\cos\theta - y\sin\theta)\,dx\,dy \tag{4.2}$$

where $g(t,\theta)$ is the projection of the object $f(x,y)$ at angle θ, and distance along a linear aperture, t. The Fourier Slice Theorem states that the 1D Fourier transform with respect to t of g is equal to the 2D Fourier transform of $f(x,y)$ along a line in 2D Fourier space tilted at the angle θ with respect to the horizontal axis. The two dimensional spatial Fourier transform of f is defined as:

$$F(k_x,k_y) = \int\int f(x,y)e^{-i(k_x x + k_y y)}\,dx\,dy. \tag{4.3}$$

There are a number of ways of proving the Fourier Slice Theorem some of which make use of certain rotational properties of the 2D Fourier transform [12, page 157] and others which rely on a change of variables [47, Section 3.2]. Our approach here is based on (4.2) directly along with the sifting property of δ functions. Taking the 1D Fourier transform of $g(t,\theta)$ with respect to t yields

$$\begin{aligned}
G(\omega,\theta) &= \int dt \left[\int\int dx\,dy\, f(x,y)\delta(t - x\cos\theta - y\sin\theta)\right] e^{-i\omega t} && (4.4)\\
&= \int\int dx\,dy\, f(x,y)\left[\int dt\,\delta(t - x\cos\theta - y\sin\theta)e^{-i\omega t}\right] && (4.5)\\
&= \int\int dx\,dy\, f(x,y)e^{-i[\omega\cos\theta\, x + \omega\sin\theta\, y]} && (4.6)\\
&= F(k_x,k_y)|_{k_x=\omega\cos\theta, k_y=\omega\sin\theta} = F(\omega\cos\theta,\omega\sin\theta). && (4.7)
\end{aligned}$$

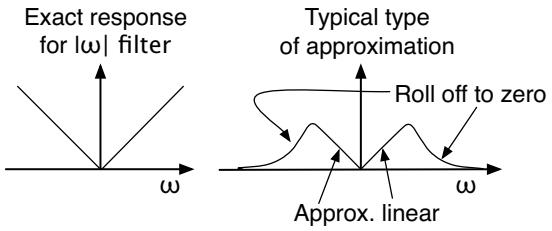


Figure 4.3: Filters used for X-ray tomography

As shown in Fig. 4.2, evaluation of $G(\omega, t)$ by allowing the Fourier variable ω run from $-\infty$ to $+\infty$ is equivalent to proceeding along the line though the $k_x - k_y$ origin in 2D Fourier space canted θ radians from the k_x axis. Moreover, as we collect projections by varying θ, we acquire spatial Fourier information about f in polar coordinates. Thus to recovery of $f(x, y)$ from projection data amounts to inverting a 2D Fourier transform in polar coordinates. The resulting technique is known as the filtered backprojection (FBP) or convolution backprojection (CBP) algorithm.

To derive FBP, we begin with the two dimensional inverse Fourier transform

$$f(x, y) = \frac{1}{(2\pi)^2} \int \int F(k_x, k_y) e^{i(k_x x + k_y y)} dk_x \, dk_y. \tag{4.8}$$

Now change variable from Cartesian coordinates, (k_x, k_y) to polar coordinates, (ω, θ):

$$k_x = \omega \cos\theta \qquad k_y = \omega \sin\theta \tag{4.9}$$

to arrive at

$$f(x, y) = \frac{1}{(2\pi)^2} \int_0^{2\pi} \int_0^{\infty} F(\omega, \theta) \omega e^{i\omega[x\cos\theta + y\sin\theta]} d\theta d\omega \tag{4.10}$$

Next, we make use of the easily proved identity [1] $F(\omega, \theta + \pi) = F(-\omega, \theta)$. Using this fact and recalling that the definition of $t = x\cos\theta + y\sin\theta$ allows us to write (4.10) as

$$f(x, y) = \frac{1}{(2\pi)^2} \int_0^{\pi} Q(t, \theta) d\theta = \int Q(x\cos\theta + y\sin\theta) \, d\theta \tag{4.11}$$

where Q is

$$Q(t, \theta) = \int_{-\infty}^{\infty} G(\omega, \theta) |\omega| e^{i\omega t} d\omega \tag{4.12}$$

Thus, the recovery of f from the projections $g(t, \theta)$ proceeds in two steps. First, we form Q according to (4.12). Second, we integrate these Q in a manner dictated by (4.11). Each of these steps deserves further discussion.

Eq. (4.12) is nothing more than an inverse Fourier transform of a product in the Fourier domain between the transformed projection and the function $|\omega|$. Hence for a given θ, $Q(t, \theta)$ is a filtered version of the corresponding projection where the frequency response of the filter is the absolute

[1] EXERCISE: Prove it

value of ω. Recall that the Fourier transform of the derivative of the signal is $i\omega$ times the transform of the signal itself. Thus the $|\omega|$ response seen here amounts to the differentiation of each projection. The frequency response of such a filter grows without bound as ω increases. Thus in practice, noise in the data at high frequencies would tend to be amplified by the filter. As illustrated in Fig. 4.3, to counter these effects one typically implements an approximation to $|\omega|$ which is linear over the band where useful data is to be found and roles off to zero at high frequencies. Details concerning the construction of such filters can be found in [47, Chapter 3].

The formation of f from the filtered projections is achieved through the integration in (4.11). This process, commonly called *backprojection*, involves "smearing" each filtered projection into the image plane and then adding the results angle by angle. This last step of addition is just the integral over angle in (4.11). The smearing interpretation is illustrated in Fig. 4.4. For a given θ, the integrand in (4.11) is $Q(t, \theta)$ evaluated at the point $t = x \cos\theta + y \sin\theta$. In the $x - y$ plane, we know from § 3.2, that this relationship among t, θ, x and y describes a line offset by t from the origin and at an angle $\pi/2 + \theta$ from the x axis. So, (4.12) indicates that we assign all points on this line the value $Q(t, \theta)$. Vary t for a fixed θ then creates a "smear" of $Q(t, \theta)$ along a region of space angles $\pi/2 + \theta$ from the x axis.[2]

4.2 Diffraction Tomography: Inverting the Born Approximation

Many of the ideas used in inverting the Radon transform are evident even for the more complicated physical problem of solving the linearized inverse scattering problem, *i.e.* developing a diffraction tomography imaging scheme to parallel the X-ray tomography case. Recalling the discussion in § 3.3.3, the problem we face is recovery of $f(\mathbf{r}) = k_s^2(\mathbf{r})$, the perturbation in the wavenumber from the nominal background, given observations of the scattered field $\phi_s(\mathbf{r})$. Under the Born approximation, the two are linked via the linear model:

$$\phi_s(\mathbf{r}) = \int g(\mathbf{r}, \mathbf{r}')\phi_b(\mathbf{r}')f(\mathbf{r}')d\mathbf{r}'. \tag{4.13}$$

with $g(\mathbf{r}, \mathbf{r}')$ the Green's function for the problem while ϕ_b is the background field; that is the field when $f(\mathbf{r}) = 0$.

Developing an analytical inverse formula for (4.13) is by no means trivial. The precise structure depends quite heavily on the specifics of the problem including

- Is the imaging problem two dimensional or three?
- The geometry of the problem. What is the structure of the background? Is it homogeneous? Is it layered? Are there boundaries? These issues all impact quite dramatically the analytical structure of the Green's function as well as the form of the incident field.
- The type of sources being used to create the incident field. In some cases, plane waves are employed. In other instances, point sources are used. To apply these methods to real data, the radiation patterns of the transmitters and receivers need to be taken into account either though some sort of calibration stage or directly in the modeling.

[2]Need to add more material here. Some ideas: fan beam, cone beam, parallel beam in 3D, implementation issues. Also, a picture with $f(x, y)$, $F(k_x, k_y)$, $Q(t, \theta)$ and the final reconstruction.

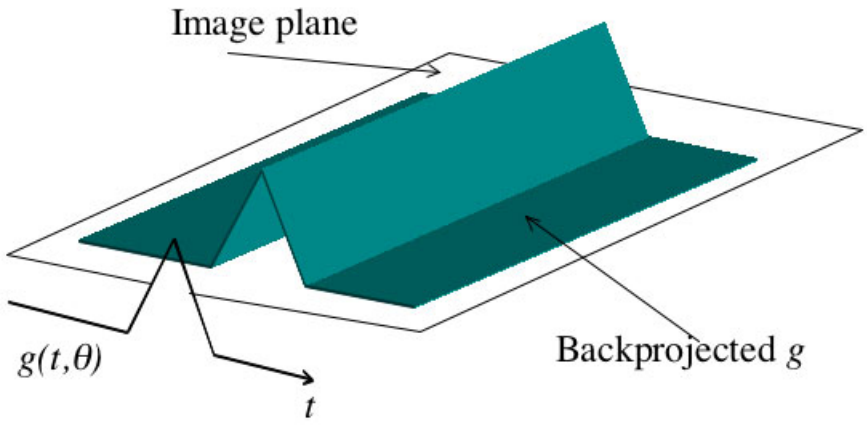


Figure 4.4: Graphical illustration of backprojection operation. The values of the projection $g(t, \theta)$ are smeared back into the image plane along lines of constant t.

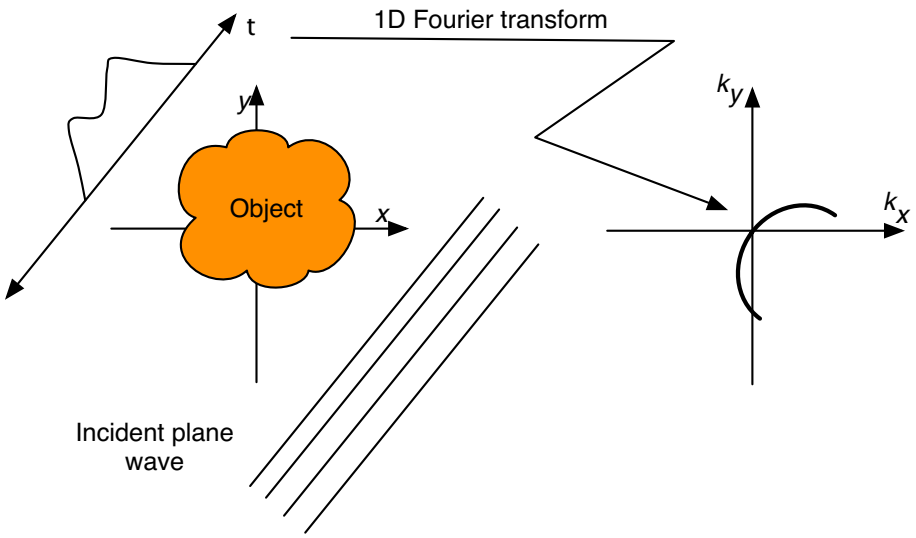


Figure 4.5: Diffraction tomography

- The nature of the medium. Is there loss or attenuation? Is it dispersive? Is it isotropic?

Variations in these three issues lead to different inverse schemes. It is not our intention here to exhaustively discuss all of these possibilities. Rather, we consider the most basic problem and leave the reader to explore the literature for variations and extensions. To be more specific, we are interested in two dimensional imaging using plane wave excitation for a homogeneous, loss-free isotropic medium. As shown in Fig. 4.5, we assume that the data are collected along linear apertures oriented parallel to the plane wave fronts of constant phase. While not necessarily the most realistic setting, this one is the closest to that of the X-ray tomography problem we just discussed. Thus, the parallels between the two are easily seen thereby providing the reader with a strong basis for examining more complex (and realistic) problems.

Given these choices, the specifics of the problem are as follows. First, the choice of an incident field generated by a plane wave yields a background field of the form

$$\phi_b(\mathbf{r}) = e^{-i\mathbf{k}\cdot\mathbf{r}} = e^{-i(k_x x + k_y y)} \tag{4.14}$$

where k_x and k_y are the components of the wavevector $\mathbf{k}$. By substituting (4.14) into the Helmholtz equation

$$\nabla^2\phi_b(\mathbf{r}) + k_b^2\phi(\mathbf{r}) = 0$$

and performing a bit of calculus and algebra[3], we see that (4.14) is a solution if and only if the dispersion relation

$$k_x^2 + k_y^2 = k_0^2 \tag{4.15}$$

is satisfied by k_x and k_y. For our problem, k_0^2 is equal just $2\pi\omega/c$ where ω is the temporal frequency

[3]EXERCISE: Do the math

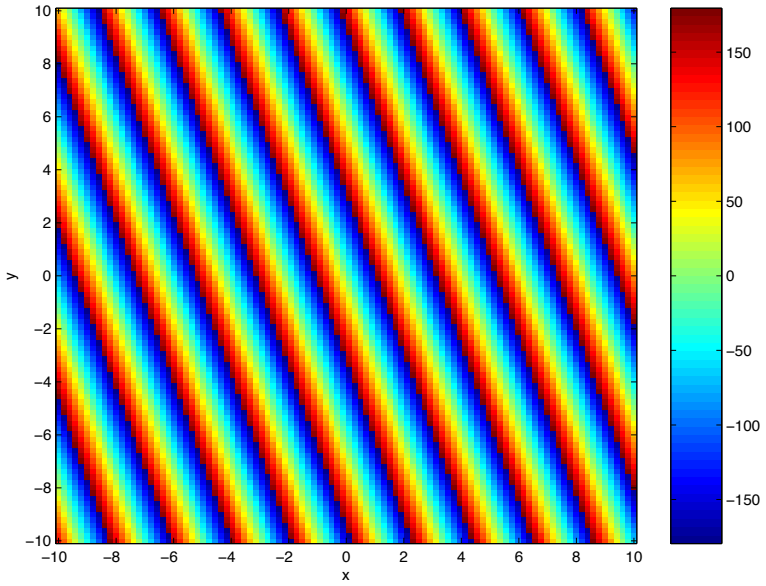


Figure 4.6: A plot of the phase of a plane wave with $k_x = 3$ and $k_y = 1$.

and c the speed of propagation in the medium. Assuming the dispersion relation is satisfied, the resulting field is everywhere constant amplitude with phase fronts that are linear functions of space. To see this we note that (4.14) implies constant phase for all values of x and y for which $k_x x + k_y y$ is constant. As illustrated in Fig. 4.6, this is just a line in $x - y$ space whose slope is determined by $-k_x/k_y$. The second implication of our choice of problem structure is the Green's function. In 2D $g(\mathbf{r}, \mathbf{r}')$ is not given by (3.19), but rather

$$g(\mathbf{r}, \mathbf{r}') = \frac{-i}{4} H_0^{(1)}(k_0|\mathbf{r} - \mathbf{r}'|) \tag{4.16}$$

where $H_0^{(1)}$ is the first kind Hankel function of order zero, [2, Chapter 7]. While this function may not be immediately familiar to the reader, its behavior is quite similar to that of the three dimensional Green's function in (3.19). Defining the scalar $\rho = |\mathbf{r} - \mathbf{r}'|$, shown in Fig. 4.7 are plots of the real and imaginary parts as well as magnitudes and phases of

$$\frac{e^{ik_0\rho}}{\rho} \qquad \text{and} \qquad H_0^{(1)}(k_0\rho)$$

function for $c = 1$ and $\omega = 3$. We see both oscillate at a spatial frequency equal to $\lambda = 2\pi/k_0 = 1/3$ and have a linear phase structure. It is shown in *e.g.* [50] that the asymptotic form of the Hankel function as $\rho \to \infty$ is

$$H_0^{(1)}(x) \approx \sqrt{\frac{2}{\pi x}} e^{i(x-\pi/4)}$$

quite similar to (3.19) except for the slightly slower rate of decay in magnitude and the phase shift. Thus while the *specifics* differ between two dimensions and three, the basic physics are qualitatively

quite similar so that intuition and understanding developed for the simpler 2D problem will carry over quite nicely when examining more complicated cases.

Just as the basis for filtered backprojection is the Fourier-Slice theorem, the inverse method we develop here, known as Fourier Backpropagation, rests on close relationship between the one dimensional Fourier transform of the data along an aperture and the two dimensional transform of the object. Specifically, as indicated in Fig. 4.5, rather than evaluating the 2D transform along a line in $k_x - k_y$ space, the contour is a semi-circle through the origin. The radius of the semi-circle is equal to $k_0^2 = \frac{2\pi\omega^2}{c^2}$ and the orientation of the circle depends on the angle of incidence of the plane wave relative to the x axis.

This result implies that the information conveyed by the data regarding the structure of the object is inherently limited in a linear inverse scattering problem. Suppose that we are able to acquire data at all angles around an object. In the case of the Radon transform problem the Fourier slice theorem indicates that this level of angular diversity in the measurements yields information about f over the whole spatial Fourier plane. That is, in the absence of noise, the object can be exactly recovered from full X ray tomography data. This is not the case for linearized inverse scattering. Indeed, assuming we probe with a single frequency, the union of the semi-circular arc in Fourier space is a circle of radius $\sqrt{2}k_0^2$. Thus in this ideal case, the data provide information only about the low spatial frequencies of the object. Thus without the use of *a priori* information as part of the reconstruction method (see Chapters 5 and 6), one can only hope to obtain a bandlimited representation of the unknown.

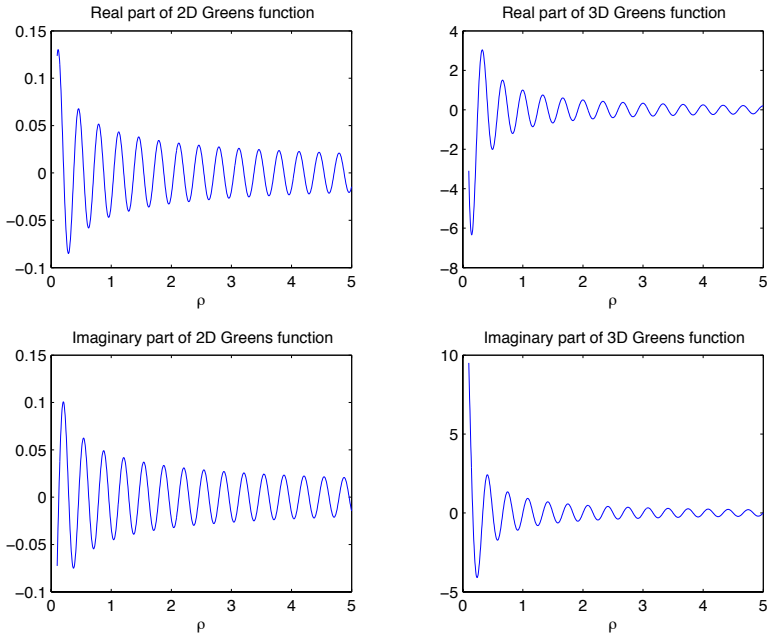
Because k_0 is proportional to the frequency, ω, we see that larger bandwidth can be obtained as we probe with higher frequencies. In other words, the ability to resolve high spatial frequency structure in an object requires correspondingly high temporal probing frequencies. Such a result should not be surprising in light of well known resolution theorae in optics which state that the ability to unambiguously distinguish closely spaced point-like objects requires the probing radiation have a frequency inversely proportional to the spacing [10][4]. Moreover because the semi-circles become lines though the origin as $\omega \to \infty$ we recover X-ray tomography as a high frequency limit of diffraction tomography.

To derive the Fourier-Diffraction theorem requires a spatial frequency domain decomposition of the incident field and the Green's function. That is, referring to (4.14), we need to be able to represent these quantities as superpositions of plane waves. Clearly, this is already the case for the incident field. For the Green's function, such a plane wave decomposition is not particularly easy to derive [18, Section 2.2], but certainly does exist and is given as

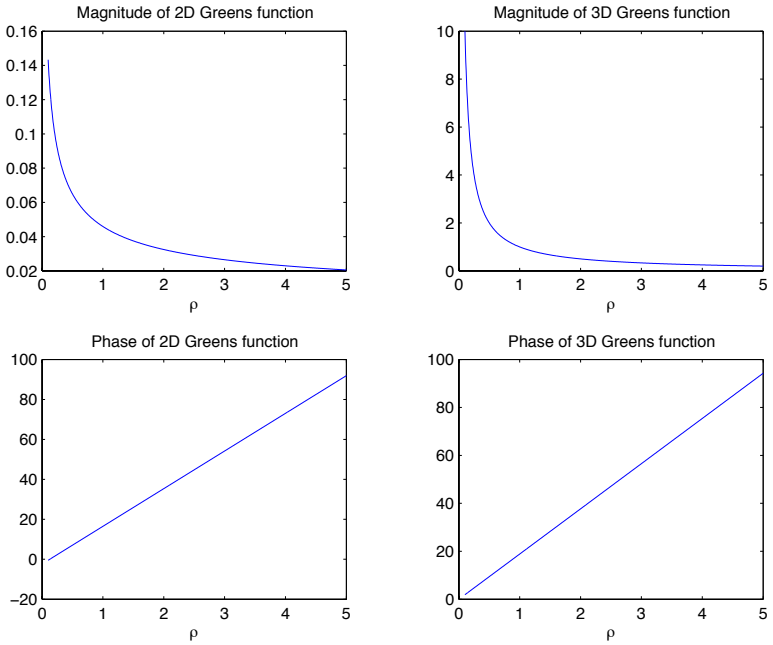
$$H_0^{(1)}\left(k_0\left|\mathbf{r}-\mathbf{r}'\right|\right) = \frac{1}{\pi}\int_{-\infty}^{\infty}\frac{1}{\sqrt{k_0^2-\kappa^2}}e^{i\left(\kappa(x-x')+\sqrt{k_0^2-\kappa^2}|y-y'|\right)}d\kappa \tag{4.17}$$

where $\mathbf{r} = [x\ y]^T$ and similarly for $\mathbf{r}'$. For notational simplicity it is common to define $\gamma = \sqrt{k_0^2 - \kappa^2}$. This equation essentially states that we can construct the Hankel function as a weighted sum of two dimensional plane waves whose spectral components exist in the spatial Fourier domain at very specific locations $(k_x, k_y) = (\kappa, \sqrt{k_0^2 - \kappa^2}) = (\kappa, \gamma)$. As κ proceeds from $-\infty$ to $+\infty$, γ is imaginary for $|\kappa| > k_0$ and real for $|\kappa| \leq k_0$. For real the (κ, γ) traces out a semi-circle in $k_x - k_y$

[4]EXERCISE: Rayleigh diffraction limit



(a) Real and imaginary parts of 2D and 3D Green's functions



(b) Magnitude and phase plots of 2D and 3D Green's function

Figure 4.7: Structure of Green's functions in two and three dimensions

space. The wavevector $\mathbf{k}$ for these *propagating* waves is purely real. Alternatively, when γ is imaginary, the wavevector acquires an imaginary k_y component and plane waves in (4.17) possess and exponential decay in y. Because such *evanescent* waves have negligible impact for problems where the sources and receivers are more than a couple of wavelengths separated (so-called *far field* imaging problems), they were initially ignored when deriving diffraction tomography [23]. However many applications arising in the twenty year since DT was discovered are not characterize by such length scales. Such problems are termed *near field* in recognition of the fact that the scattering is taking place within a couple of wavelengths of the sensors. Thus, there has been considerable interest in this period of time in the development of tomographic inversion formulae which make use of these decaying waves.

To use (4.17) in relating the 1D Fourier transform of the data to the 2D transform of the object, let us assume for a moment that we illuminate with a plane wave traveling in the $+y$ direction and measure the scattered field along the line $y = y_0$ where y_0 is sufficiently large that it does not intersect the region to be imaged. In this case, the incident field is $e^{ik_0 y}$. Using this fact and (4.17) in (4.13) yields

$$\phi_s(x, y = l_0) = -\frac{i}{4\pi} \int_{-\infty}^{\infty} \frac{d\kappa}{\gamma} \int f(x', y') e^{i(\kappa(x-x') + \gamma(l_0 - y'))} e^{ik_0 y'} \, dx' \, dy'. \tag{4.18}$$

By recognizing

$$\int f(x', y') e^{-i(\kappa x' + (\gamma - k_0) y')} \, dx' \, dy'$$

as $F(\kappa, \gamma - k_0)$, the two dimensional Fourier transform of f evaluated at $k_x = \kappa$ and $k_y = \gamma - k_0$, we see that (4.18) can be written as

$$\phi_s(x, y = l_0) = -\frac{i}{4\pi} \int_{-\infty}^{\infty} \frac{1}{\gamma} e^{i(\kappa x + \gamma l_0)} F(\kappa, \gamma - k_0) d\kappa. \tag{4.19}$$

Now, we evaluate the 1D Fourier transform of ϕ_s with respect to t, that is, along the line where the fields are observed. By definition we have

$$\Phi_s(\omega, l_0) = \int \phi_s(t, l_0) e^{-i\omega t} \, dt.$$

The transform of the right hand side of (4.19) with respect to x requires only the evaluation of $\int e^{i(\kappa - \omega)x} dx$ which is $2\pi\delta(\omega - \kappa)$. Hence after a bit of algebra, we conclude

$$\Phi_s(\kappa, l_0) = -\frac{i}{2\sqrt{k_0^2 - \kappa^2}} e^{il_0\sqrt{k_0^2 - \kappa^2}} F(\kappa, \sqrt{k_0^2 - \kappa^2} - \kappa) \tag{4.20}$$

which is the desired result. Specifically, the transform of the data when viewed as a function of κ traces out the locus of points $k_x = \kappa$, $k_y = \sqrt{k_0^2 - \kappa^2} - \kappa$ in the (k_x, k_y) plane. Imposing the condition that $|\kappa| < k_0$ so that the square root remains real-values, this set of points corresponds to half of the circle defined by the equation

$$k_x^2 + (k_y + k_0)^2 = k_0^2;$$

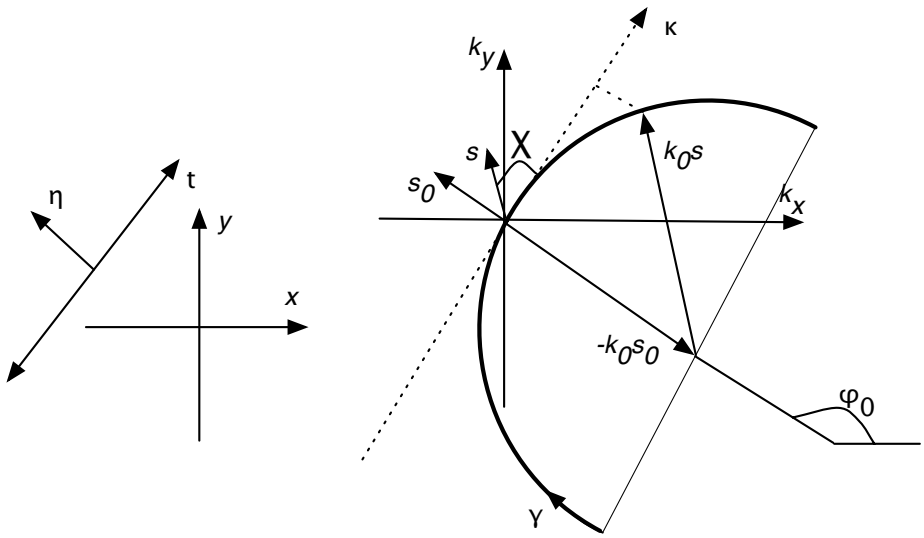


Figure 4.8: Coordinate Systems for Filtered Backpropagation

a circle of radius k_0 that passes though the origin. While this derivation was based on the assumption that the incident wave was propagating in the $+y$ direction, physically, there is nothing preferred in the setup of the problem to this orientation. By rotating the direction of incidence, one would obtain a rotated semi-circular contour. Mathematically demonstrating this is the subject of an exercise at the end of this chapter.[5]

Fig. 4.8 provides another manner of parameterizing this contour which will be of use in subsequent discussion. Any point on the semi-circle can be represented as the vector $k_0(\mathbf{s} - \mathbf{s}_0)$. The quantity $\mathbf{s}_0$ is a unit vector in the spatial Fourier domain pointing in the same direction as the incident field, φ_0. The unit vector $\mathbf{s}$ on the other hand is at an angle χ with respect to the κ axis. As χ changes, we sweep out the points on the desired semi-circular arc. By the Fourier-Diffraction theorem, these points are directly related to the receiver line over which the scattered data are collected.

These geometric ideas are central in deriving the Filtered Backpropagation algorithm to obtain a direct reconstruction formula for the recovery of f from the Born-type data. While the derivation of the algorithm is a bit tedious, it is also instructive as many of the steps represent generalizations of those encountered in developing the Filtered Backprojection methods in § 4.1. We begin with the inverse Fourier transform for $f(x, y)$,

$$f(x,y) = \frac{1}{(2\pi)^2} \int dk_x \int dk_y F(k_x, k_y) e^{i(k_x x + k_y y)}. \tag{4.21}$$

A sequence of two changes of variable are made to transform the integration from k_x and k_y to variables more relevant to the underlying problem. Specifically, the ultimate goal is a double

[5]EXERCISE: rotating the FDT.

integral over all incidence angles, φ_0, and then along κ, the Fourier variable conjugate to t. As shown in Fig. 4.8, there is a one-to-one correspondence between κ and points on the semi-circle at angle φ_0. Hence, by using κ as a variable of integration, we shall be able to obtain an inversion scheme that explicitly accesses the data in the format that they are provided by the physics of the problem. The changes of variable which achieve this are as follows:

$$\begin{bmatrix} k_x \\ k_y \end{bmatrix} \rightarrow \begin{bmatrix} \chi \\ \phi_0 \end{bmatrix} \rightarrow \begin{bmatrix} \kappa \\ \phi_0 \end{bmatrix} \tag{4.22}$$

via

$$\begin{bmatrix} k_x \\ k_y \end{bmatrix} = k_0(\mathbf{s} - \mathbf{s}_0) \quad \text{with} \quad \mathbf{s} = \begin{bmatrix} \sin\varphi_0 \\ \cos\varphi_0 \end{bmatrix} \quad \mathbf{s}_0 = \begin{bmatrix} \sin\chi \\ \cos\chi \end{bmatrix} \tag{4.23}$$

and then

$$\cos\chi = \frac{\kappa}{k_0} \qquad \sin\chi = \frac{\gamma}{k_0} \quad \gamma = \sqrt{k_0^2 - \kappa^2} \tag{4.24}$$

Upon making these changes and being careful with the required Jacobian calculations we arrive at

$$f(x,y) = \frac{1}{2\pi}\frac{k_0}{2}\int_{-\pi}^{\pi} d\varphi_0 \int_{-k_0}^{k_0} \frac{d\kappa}{\gamma}|\kappa| F\left(k_0(\mathbf{s}-\mathbf{s}_0)\right) e^{ik(\mathbf{s}-\mathbf{s}_0)}. \tag{4.25}$$

Now we make use of the following two facts:

1. Under the Born approximation the Fourier diffraction theorem, (4.20) states

$$F(k_0(\mathbf{s}-\mathbf{s}_0)) = i2\gamma\,\Phi_s(\kappa, l_0)e^{-il_0\gamma}$$

2. If we use the $\eta - t$ coordinate system in Fig. 4.8, then

$$k_0(\mathbf{s}-\mathbf{s}_0) = \kappa t + (\gamma - k_0)\eta$$

to express (4.25) as

$$f(x,y) = \frac{1}{(2\pi)^2}\frac{k_0}{2}\int_{-\pi}^{\pi} d\varphi_0 \int_{-\infty}^{\infty} \left[|\kappa|\Phi_s(\kappa, l_0)e^{-il_0\gamma}B_\eta(\kappa)\right] e^{i\kappa t}\, d\kappa. \tag{4.26}$$

where

$$B(\kappa) = \begin{cases} e^{i[(\gamma-k_0)\eta]} & |\kappa| < k_0 \\ 0 & \text{else} \end{cases} \tag{4.27}$$

To clarify the connections among (4.26), the notion of backpropagation, and filtered backprojection, we define the inner integral as $\Pi(t,\eta)$. Noting from Fig. 4.8 that t and η are related to x and y via

$$\begin{aligned} t &= x\sin\varphi_0 - y\cos\varphi_0 \\ \eta &= x\cos\varphi_0 + y\sin\varphi_0 \end{aligned}$$

allows us to write (4.26) as

$$f(x,y) = \frac{1}{(2\pi)^2}\frac{k_0}{2}\int_{-\pi}^{\pi} d\varphi_0 \Pi(x\sin\varphi_0 - y\cos\varphi_0, x\cos\varphi_0 + y\sin\varphi_0) \tag{4.28}$$

Equation (4.28) is basically equivalent to (4.12) in the case of inverting the Radon transform in that both implement the sum over incident angles of frequency-domain processed forms of the measured fields. The manner in which the data are processed is also quite similar between X-ray and diffraction tomography in the both make use of the $|\omega|$-type of filter. Aside from the trivial scaling of the data by $e^{il_0\gamma}$, the real source of difference between inverting the Radon transform and diffraction tomography lies in the presence of the B_η filter. To fully appreciate the physical significance of $B_\eta(\kappa)$ requires a small, but interesting digression into diffractions theory and the notion of an angular spectrum of waves.

Consider the problem of determining the field that satisfies the Helmholtz equation in free space in the halfspace $y > 0$ given knowledge of the field on the plane $y = 0$. In other words, given $\phi(0, y)$ we seek a $\phi(x, y)$ satisfying

$$\frac{\partial^2}{\partial x^2}\phi(x,y) + \frac{\partial^2}{\partial y^2}\phi(x,y) + k_0^2\phi(x,y) = 0.$$

Defining the y dependent Fourier transform of ϕ as

$$A(\kappa, y) = \int \phi(x,y)e^{-i\kappa x}\,dx \tag{4.29}$$

allows us to write the Helmholtz equation in terms of A

$$\frac{\partial}{\partial y^2}A(\kappa,y) + \gamma^2 A(u,y) = 0 \tag{4.30}$$

where $\gamma = \sqrt{k_0^2 - \kappa^2}$. The solution of (4.30) is

$$A(\kappa,y) = C_+(u,0)e^{i\gamma y} + C_-(u,0)e^{-i\gamma y}. \tag{4.31}$$

where the first term represents a plane wave traveling in the $+y$ direction (*i.e.* $\mathbf{k}$ for this wave is $[0\ \gamma]^T = [0\ \sqrt{k_0^2 - \kappa^2}]^T$) and the second is a plane wave traveling in the $-y$ direction. Because the source of the field is assumed to lie in the region $y < 0$, $C_-(u, 0) = 0$. Thus

$$A(\kappa,y) = C_+(u,0)e^{iy\sqrt{k_0^2-\kappa^2}} \tag{4.32}$$

and using the inverse (4.29) we conclude

$$\phi(x,y) = \frac{1}{2\pi}\int C_+(u,0)e^{i(ux+\kappa y)}\,du \tag{4.33}$$

From (4.32) and (4.33) we draw a number of conclusions:

1. By evaluating (4.33) at $y = 0$ and taking an inverse Fourier transform we see that $C_+(u, 0) = A(u, 0)$ the Fourier transform of $\phi(x, 0)$, the data from which we want to compute the fields for $y > 0$.

2. Comparing (4.14) and (4.33) indicates that (4.33) is a representation of $\phi(x, y)$ as a superposition of plane waves where $k_x = u$ and $k_y = \kappa = \sqrt{k_0^2 - u^2}$ and the plane wave spectrum is given by $A(u, 0)$.

3. Taking a systems view of the propagation problem, we can view the input as $\phi(x, 0)$ and the output as $\phi(x, y)$. According to (4.32), the transfer function of this system is

$$H(\kappa) = \frac{A(\kappa, 0)}{A(\kappa, y)} = e^{iy\sqrt{k_0^2 - \kappa^2}} \tag{4.34}$$

   For $|\kappa| < k_0$, we have propagating fields and the transfer function is merely a y-dependent phase shift to the initial angular spectrum. For the case of evanescent waves where $|\kappa| > k_0$, it is not hard to show that the transfer function decays exponentially fast as a function of distance from the plane $y = 0$. Finally because the physical significance of H is to move the fields from the boundary $y = 0$ into the space $y > 0$, we call H the *propagator*.

The connection to (4.26) and (4.27) should now be clearer. The quantity $B_\eta(\kappa)$ encountered in the processing of inverting the Born approximation is basically a propagator from the aperture where the data are recorded into the region to be imaged. More technically, it is a bandlimited version of the adjoint of the propagator. As explained in [23], when ignoring the evanescent fields, the two operators are equivalent.
[6]

4.3 Exercises

4.1 In this problem, we examine the filtered backprojection (FBP) approach and write MATLAB routines for the reconstruction of sampled data. The FBP algorithm is a two step process for recovering an image from its projections. It represents an exact, closed form inverse of the continuous Radon transform. The FBP method is mathematically represented as

$$q(\theta, t) = \int_{-\infty}^{\infty} g(\theta, \omega)|\omega| e^{j2\pi\omega t} d\omega \tag{4.35}$$

$$f(x, y) = \int_0^{\pi} q(\theta, x\cos\theta + y\sin\theta) d\theta \equiv (\mathbf{B}q)(x, y) \tag{4.36}$$

where $q(\theta, t)$ is the filtered projection at angle θ, $g(\theta, \omega)$ is Fourier transform of the projection at angle θ taken with respect to the t coordinate, and $\mathbf{B}$ is the so-called backprojection operator.

1. Please provide a graphical and descriptive interpretation of (4.36). Specifically, justify the use of the term "backprojection."

[6] It might be nice to add the work of John Schotland here as a segue into the numeric approach to the problem

2. The time domain representation of the filter $|\omega|$ does not exist as a well behaved function. Explain why this is so.

3. Typically, one assumes that the projections are band-limited to $\omega \in [\Omega, \Omega]$. Under this assumption it is natural to define a "windowed" form of the $|\omega|$ filter by multiplying this filter with a box function in the frequency domain. Now the filter has a tractable time domain representation. Find it and use MATLAB to examine its characteristics as Ω is increased.

4. Starting from (4.35), show that one can write $q(\theta, t)$ as

$$q(\theta, t) = h(t) * \frac{\partial g(\theta, t)}{\partial t}. \tag{4.37}$$

   where $*$ is convolution. Specifically, find $h(t)$. (Hint: This $h(t)$ also plays a role in demodulation for analog communication systems.) Comment on the relationship between (4.37) and the filter discussed in (b.i). In particular, it would appear that we have here the filter which previously we said did not really exist. What's up?

5. We will now build on what you know to create an implementation of the FBP reconstruction for the discrete case. In particular, write a MATLAB routine called **fbp.m** that takes as input the discrete data vector y, the corresponding projection matrix T, and the number of angles N_θ used in generating y and produces the FBP reconstruction $\widehat{x}$ as output. Notes: 1) MATLAB's **fft** routine places the frequency origin at the first sample, thus depending on how you create the $|\omega|$ filter you may need to use the **fftshift** routine. 2) Since T is being passed in, the implementation of the backprojection operation is very simple and should not require extensive coding. 3) Do not worry about the global scaling of the reconstruction.

6. Create an approximation to an impulse object f_o by making a 32×32 image which is all zeros except for a single 2×2 region of ones near the center (you might want to use **makeb.m** to do this). Generate FBP reconstructions of this object from projection data corresponding to N uniformly spaced projection angles in [0, 180), and N projections per angle for $N \in \{16, 32, 64, 128\}$. For each N, plot the central cross-section through the corresponding reconstruction.Based on this experiment and your own visual evaluation, what is the ratio of the number of observations to the number of unknowns (i.e. pixels) that is necessary before FBP produces reasonable reconstructions in the discrete case?

7. From the class web site retrieve the file **phantom.mat.gz** containing a 32×32 phantom. Generate the FBP reconstructions for $N \in \{16, 32, 64, 128\}$ angles in $[0, 180)$. Based on this experiment and your own visual evaluation, what is the ratio of the number of observations to the number of unknowns (i.e. pixels) that is necessary before FBP produces reasonable reconstructions in the discrete case?

Chapter 5

Numerical Methods for Linear Inverse Problems

Methods such as filtered backprojection and filtered backpropagation which we examined in the last chapter form the basis for widely used methods in a range of application areas including medical imaging with CAT and MRI as well as geophysical imaging using acoustics and electromagnetics. In other areas however the problems are such that these and related methods are either not appropriate or do not provide the performance required for the application. For example, filtered backprojection was derived specifically for tomographic problems where the Fourier-slice theorem can be shown to hold while filtered backpropagation was of use in cases where the Fourier diffraction theorem governed the relationship between the data and the object to be recovered. If the physics of the sensor do not match those for which the algorithm was derived (*e.g.* if the Born approximation is not accurate), then one would expect a decline in performance as evidenced by poorer imagery.

Alternatively and quite typically, the quantity of data available to an inversion method will impact the choice and utility of an algorithm. The filtered backprojection algorithm is ideal when one possess data for a dense collection of angles θ between 0 and π. In many circumstances the angles over which data can be acquired are severely limited. Such limited view problems arise for example in the cases of synthetic aperture imaging [71,83], medical imaging [1,9,33,89], geophysics [37, 53, 66, 90], and nondestructive evaluation [56, 57, 63]. While there is nothing preventing one from using a method such as filtered backprojection to process these data, the paucity of data will result in the presence of non-physical artifacts in the imagery.

LIMITED VIEW FBP EXAMPLE GOES HERE. MAYBE ALSO A DECONVOLUTUION PROBLEM FOR A LOWPASS FILTER.

To quite a large extent, the issues and problems arising in this example typify those encountered in a very broad range of inverse problems. Specifically, the data provide far from complete information regarding the object to be imaged. The lack of information may be due to limited view issues. Often however the physics of the problem inherently limits information. While we did not emphasize it at the time, this was the case for diffraction tomography wherein the data provided for the recovery of only a spatially bandlimited version of the true profile. Thus, high frequency information such as the precise location of edges, is not conveyed by the data. In many cases both the physics as well as limited sensor placement work together to severely complicate the inversion

process.

There are a range of approaches for building on the tools developed in the previous chapter to handle issues such as these. At one end of the spectrum is the development of analytic inversion methods (akin to limited view versions of the approaches in Chapter 4) appropriate for the physics of the underlying problem and the specifics of the sensor geometry. Such techniques are both interesting and useful; however they also tend to be highly specific to the application under investigation. Alternatively, one can use the tools of § 3.4 to first discretize the problem and then look at inversion more generically using ideas drawn from vector space analysis, numerical linear algebra, optimization theory, etc. The power of this later approach is the wide applicability of the resulting methods. Moreover as we shall see later in this chapter, this approach provides us with the ability to easily augment the information in the data with any prior information we may have concerning the structure of the region in order to improve the quality and usefulness of the resulting reconstruction.

The primary shortcoming of a numerical view of inversion is the loss of insight incurred when one reduces the physics of the problem to discrete form. In the case of linear inverse problems for example one typically reduces a linear integral equation to a matrix-vector problem. As we shall see, tools such as the singular value decomposition (SVD) then play a dominant role in the analysis and solution of the resulting inverse problem. Often forgotten (or at least not explicitly taken into consideration) however is the physics which underlies these problems. Thus, the analytical elegance as well as the algorithmic implications of results such as the Fourier diffraction theorem are not typically encountered in the domain of discrete inverse problems. Bridging the gap between application specific inversion methods and purely numerical approaches to their solution is an area of considerable research.

In this chapter we shall develop the tools and methods common to the more numeric approach to linear inverse problems. The majority of this chapter will be concerned with problems for which the object as well as the data are both finite dimensional and are related via a matrix-vector model of the form

$$\mathbf{g} = \mathbf{K}\mathbf{f} \tag{5.1}$$

In (5.1), $\mathbf{g} \in \mathbb{R}^M$ is the data vector $\mathbf{f} \in \mathbb{R}^N$ is a vector of unknowns to be determined, and $\mathbf{K}$ is a discretized form of the linear operator, that is, an $M \times N$ matrix, relating the two. Finally, we use $\hat{\mathbf{f}}$ to denote a reconstructed estimate of $\mathbf{f}$ obtained through the processing of a set of data $\mathbf{g}$.

As alluded to previously, problems of the form (5.1) are quite amenable to analysis and solution using standard techniques from basic linear algebra. The intuition gained from such analysis does carry over to more complex problems. Specifically, in the last part of this chapter we are concerned with semi-discrete inverse problems where again the data are discrete, but this time, we wish to recover a continuously-values object. More sophisticated Hilbert space ideas can and will be brought to bear on this class of problems, but at a fundamental level, they are for the most part strikingly similar to those used for the fully discrete case.

5.1 Ill-posedness

One typically refers to inverse problems where the information content of the data is limited due to conditions related to the physics of the sensing modality or restrictions on the placement of

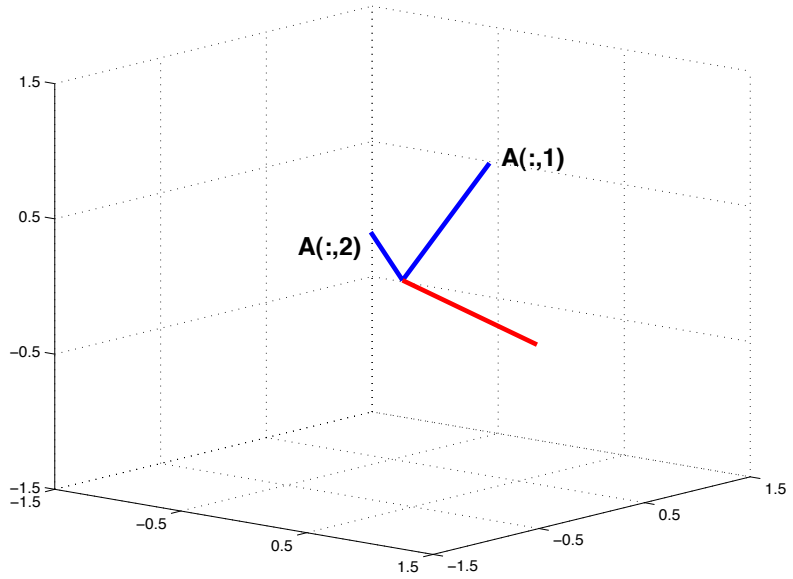


Figure 5.1: Blue lines illustarte the span of the columns of A in (5.2). The red line is the orthogonal complement of the span in $\mathbb{R}^3$

the sensors as being *ill-posed*. The precise mathematical notion of an ill-posed problem was first formulated in 1902 by the mathematician Jacques Hadamard who was studying the manner in which the solution to certain partial differential equations was dependent on the boundary data. According to Hadamard, a problem was well posed if three conditions are met:

1. At least one solution exists.
2. The solution is unique.
3. The solution is stable in the sense that its dependence on the boundary data is continuous. Less formally, but perhaps more clearly, the notion of stability implies that small changes in the boundary data should not yield overly large changes in the resulting solution to the underlying problem.

Within the context of the model inverse problem in (5.1), the first two criteria of Hadamard require the existence and uniqueness of a solution $\mathbf{f}$ for a given set of data $\mathbf{g}$. Technically, for finite dimensional problems, if a solution exists it will always be continuously dependent on $\mathbf{g}$ where continuity is formally defined in the $\delta - \epsilon$ sense of mathematical analysis. Thus, here the issue of stability will be interpreted a bit less formally using in terms of the intuitive idea that the presence of "small" perturbations in the data due for example to sensor noise, model mismatch, calibration errors etc., should not result in "large" changes to $\hat{\mathbf{f}}$.

5.1.1 Existance

The finite dimensional model provides a natural setting to make clear the three concepts of existence of a solution, uniqueness of solutions, and stability. Linear systems for which there are more rows in

K than columns, termed overdetermined systems, provide classic examples of cases where solutions may not exist. As an example, consider the problem

$$\begin{bmatrix} g_1 \\ g_2 \\ g_3 \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 3 & -4 \\ 4 & 3 \end{bmatrix} \begin{bmatrix} f_1 \\ f_2 \end{bmatrix} + \begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix} \tag{5.2}$$

In this case, the inverse problem is to determine the two elements of $\mathbf{f}$ which give rise to an arbitrary three vector $\mathbf{g}$. Basic linear algebra indicates that $\mathbf{f}$ is only able to generate vectors in $\mathbb{R}^3$ which lie in the linear span of the columns of $\mathbf{K}$; that is the range of the matrix $\mathbf{K}$. Geometrically, this range is shown in Fig. 5.1 as the plane containing the two blue lines. More specifically, any vector in $\mathbb{R}^2$ which contains a component along the red line in Fig. 5.1, *i.e.* perpendicular to this plane, cannot be exactly represented by any choice of $\mathbf{f}$. Under the rather mild assumption that there is generally some noise in the data, then in all but the most unlikely of circumstances, $\mathbf{g}$ will have a "red" piece. Hence a solution to (5.2) will not exist. To put it another way, in general two degrees of freedom as represented by f_1 and f_2 are generally insufficient to provide a representation for the three degrees of freedom in the vector $\mathbf{g}$.

The singular value decomposition (SVD) of the matrix $\mathbf{K}$ can be of use in algebraically understanding this problem. Recall that for an $M \times N$ matrix $\mathbf{K}$ the SVD is written as $\mathbf{K} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ where $\mathbf{U}$ is orthonormal and of size $M \times M$, $\mathbf{V}$ is also orthonormal and of size $N \times N$, and $\mathbf{\Sigma}$ is $M \times N$ and is zero except for the main diagonal where the singular values are located. For the overdetermined problem then, $\mathbf{\Sigma}$ takes the form

$$\mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_1 \\ \mathbf{0} \end{bmatrix} \tag{5.3}$$

with $\mathbf{\Sigma}_1 = \text{diag}\{\sigma_1, \sigma_2, \ldots, \sigma_N\}$ and where for simplicity here we assume that all the singular values are non-zero.[1] In terms of the SVD, the inverse problem of interest is $\mathbf{g} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}\mathbf{f}$. Defining $\tilde{\mathbf{f}} = \mathbf{V}\mathbf{f}$ and $\tilde{\mathbf{g}} = \mathbf{U}^T\mathbf{g}$ we have the equivalent problem $\tilde{\mathbf{g}} = \mathbf{\Sigma}\tilde{\mathbf{f}}$. From the discussion in § 2.2.3, this is just a version of the original problem "rotated" into a coordinate system adapted to $\mathbf{K}$. In this system, the issue of existence is much clearer. Specifically, by the structure of the matrix $\mathbf{\Sigma}$, a solution will exist if and only if the last $M - N$ components of the vector $\tilde{\mathbf{g}}$ are exactly zero, a situation not likely to be encountered in the presence of noise.

5.1.2 Uniqueness

In contrast to the existence issue which is most naturally studied in terms of *over*determined linear systems, uniqueness is best understood using the example of an *under*determined system where $M < N$. Here one has not just one solution, but an infinite number of them due to the presence of a nullspace associated with $\mathbf{K}$. Again let us start with a simple example defined by the 2×3 matrix

$$\mathbf{K} = \begin{bmatrix} -1.3 & -0.5 & 1.0 \\ -1.0 & 1.0 & 0.0 \end{bmatrix} \tag{5.4}$$

[1] Soon we shall lift this assumption.

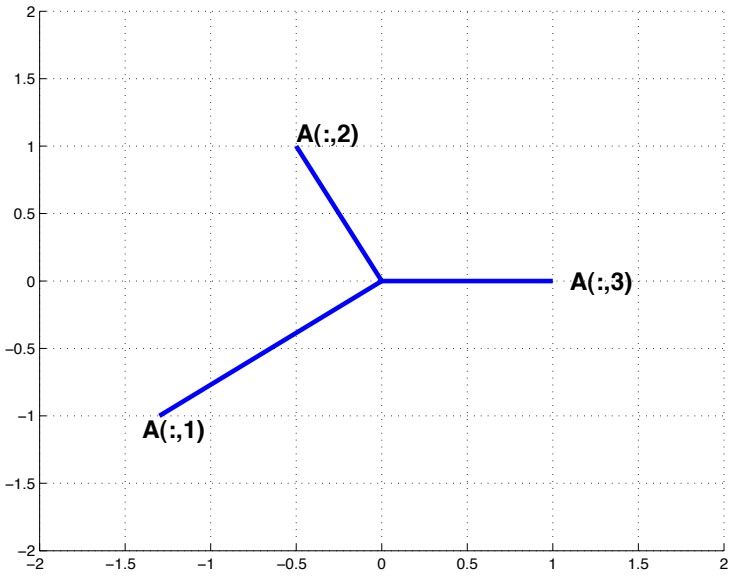


Figure 5.2: Blue lines illustarte the span of the columns of A in (5.4).

In this case the range of $\mathbf{K}$ is $\mathbb{R}^2$ and the columns are plotted in Fig. 5.2. Geometrically, the inverse problem here is to use the three degrees of freedom in $\mathbf{f}$ to build vectors in $\mathbb{R}^2$. The extra element of $\mathbf{f}$ implies that in general there will be some flexibility concerning how this is done. Indeed, this is precisely the case for the matrix $\mathbf{K}$ in (5.4). For example, say the target $\mathbf{g} = [0\ 1]^T$. Using the first two columns of $\mathbf{K}$ yields a solution $\hat{\mathbf{f}}_1 = [-0.2778\ 0.7222\ 0.0000]$. We could just as easily use the first and third columns to arrive at a solution $\hat{\mathbf{f}}_2 = [-1.0000\ 0.0000\ -1.3000]^T$. Finally, it is easy enough to verify that one (of many) solutions employing all three elements of $\mathbf{f}$ is $\hat{\mathbf{f}} = [0.1591\ 1.1591\ 0.7863]^T$.

The primary issue here is that $\mathbf{K}$ has a nullspace spanned by the vector

$$\mathbf{f}_{null} = \begin{bmatrix} 0.4369 \\ 0.4369 \\ 0.7863. \end{bmatrix}$$

Thus any solution to the problem can be written as a linear combination of a vector in $\mathcal{N}(A)^\perp$ plus a component in $\mathcal{N}(A)$. By direct calculation, one can verify that the orthogonal complement of $\mathcal{N}(A)$ is spanned by the columns of the matrix

$$\mathbf{B} = \begin{bmatrix} -0.5556 & -0.2778 \\ -0.5556 & 0.7222 \\ 0.0000 & 0.0000 \end{bmatrix} \tag{5.5}$$

so that we can write any solution to the inverse problem as $\mathbf{f} = \mathbf{B}\mathbf{f}_1 + \alpha\mathbf{f}_{null}$ where α is an arbitrary real number and $\mathbf{f}_1$ is a two-vector. This decomposition should clarify the non-uniqueness inherent in this problem. Because $\mathbf{KB}$ is just the 2×2 identity matrix and $\mathbf{K}\mathbf{f}_{null} = \mathbf{0}$, we have that

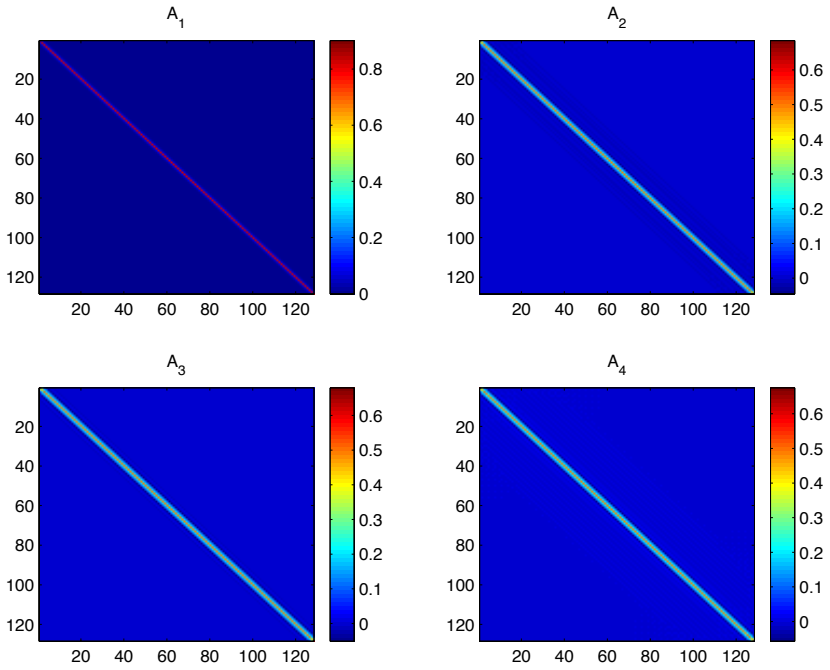


Figure 5.3: $\mathbf{K}$ matrices

$\mathbf{g} = \mathbf{f}_1$ and can therefore conclude that any solution to the inverse problem may be written as $\hat{\mathbf{f}} = \mathbf{B}\mathbf{g} + \alpha\mathbf{f}_{null}$. Hence the choice of α corresponds to the one extra degree of freedom we have in selecting a solution to the problem. Because this choice is arbitrary, there are obviously an infinite number of such solutions.

As with the overdetermined linear system, in the case where we have fewer rows than columns, the SVD provides some useful insight. Assuming that the number of nonzero singular values is now M, $\mathbf{\Sigma}$ takes the form

$$\mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_1 & \mathbf{0} \end{bmatrix} \tag{5.6}$$

with $\mathbf{\Sigma}_1 = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_M)$. Again, we define $\tilde{\mathbf{g}} = \mathbf{U}^T\mathbf{g}$ and $\tilde{\mathbf{f}} = \mathbf{V}\mathbf{f}$. If we take $\tilde{\mathbf{f}}_1$ as the first M elements of $\tilde{\mathbf{f}}$ and $\tilde{\mathbf{f}}_2$ as the remaining $N - M$ components of $\tilde{\mathbf{f}}$ then in the "rotated" domain, the linear system takes the form

$$\tilde{\mathbf{g}} = \mathbf{\Sigma}_1\tilde{\mathbf{f}}_1 + \mathbf{0}\tilde{\mathbf{f}}_2.$$

In other words as long as $\tilde{\mathbf{f}}_1 = \mathbf{\Sigma}^{-1}\tilde{\mathbf{g}}$ then $\tilde{\mathbf{f}}_2$ can be *anything* without impacting the value for $\tilde{\mathbf{g}}$. Hence the lack of uniqueness for the problem is captured explicitly by the singular value decomposition the identification of $\tilde{\mathbf{f}}_2$ as those degrees freedom whose values has no impact on the $\mathbf{K}\mathbf{f}$ product.

5.1.3 Stability

To illustrate the more subtle issue of the stability of an inverse problem, consider the problem of recovering $\mathbf{f}$ from $\mathbf{g}$ given the four $\mathbf{K}$ matrices shown in Figs. 5.3 and 5.4. Each of these matrices

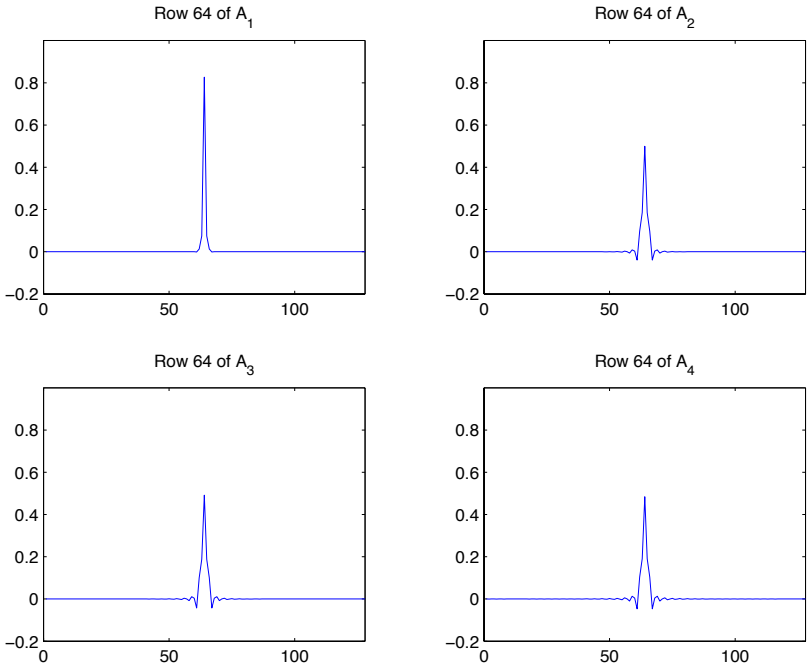


Figure 5.4: Plots of $\mathbf{K}(64,:)$ for the matrices in Fig. 5.3. Each matrix performs local averaging of the input signal.

is of size 128×128. Each of the four $\mathbf{K}$ shown in the figure have structure only near the diagonal implying that g_m is an averaged form of f_n in a narrow band of samples around $m = n$. Moreover, because the four images are more or less constant along the diagonals, the averaging kernel is not changing appreciably from one n to the next. It is only being "dragged" along. Fig. 5.4 is in fact a plot of the 64-th row of each matrix and shows that the structure of the averaging being performed by each of these different matrices is really quite similar.

In fact, these matrices have been constructed to have very specific SVD structure. All four have the same set of singular vectors and in each case $\mathbf{U} = \mathbf{V}$. Only their singular values differ. Hence for $i = 1, 2, 3, 4$ we have $\mathbf{K}_i = \mathbf{U}\mathbf{\Sigma}_i\mathbf{U}_i^T$ where $\mathbf{\Sigma}_i = \text{diag}(\boldsymbol{\sigma}_i)$ and $\boldsymbol{\sigma}_i$ is the 128×1 vector of singular values and the matrix of singular vectors $\mathbf{U} = [\mathbf{u}_1\ \mathbf{u}_2\ \ldots\ \mathbf{u}_n\ \ldots\ \mathbf{u}_{128}]$. To mimic traditional Fourier analysis, the singular vectors have been chosen to behave like sinusoids of increasing frequency as a function of n. A few are shown in Fig. 5.5. Thus, the singular values play a role quite similar to that of a traditional filter.[2] Plots of $\boldsymbol{\sigma}_i$ are shown in Fig. 5.6. Here we order the singular values not according to their magnitude but rather according to the frequency of the corresponding singular vector. In all four cases, all of the singular values are strictly greater than zero. As i varies however, we have structured the filters in such a way that they come increasingly closer to having

[2]The primary difference here is that the filter coefficients as well as the basis functions are all real valued whereas in the normal Fourier setting they are complex valued. This discrepancy is caused by the fact that complex exponentials are eigen-functions of convolution operators and hence can be complex. To be consistent with much of the rest of this manuscript, the analysis we carry out here is in terms of the SVD for which the singular values and singular vectors must be real-valued.

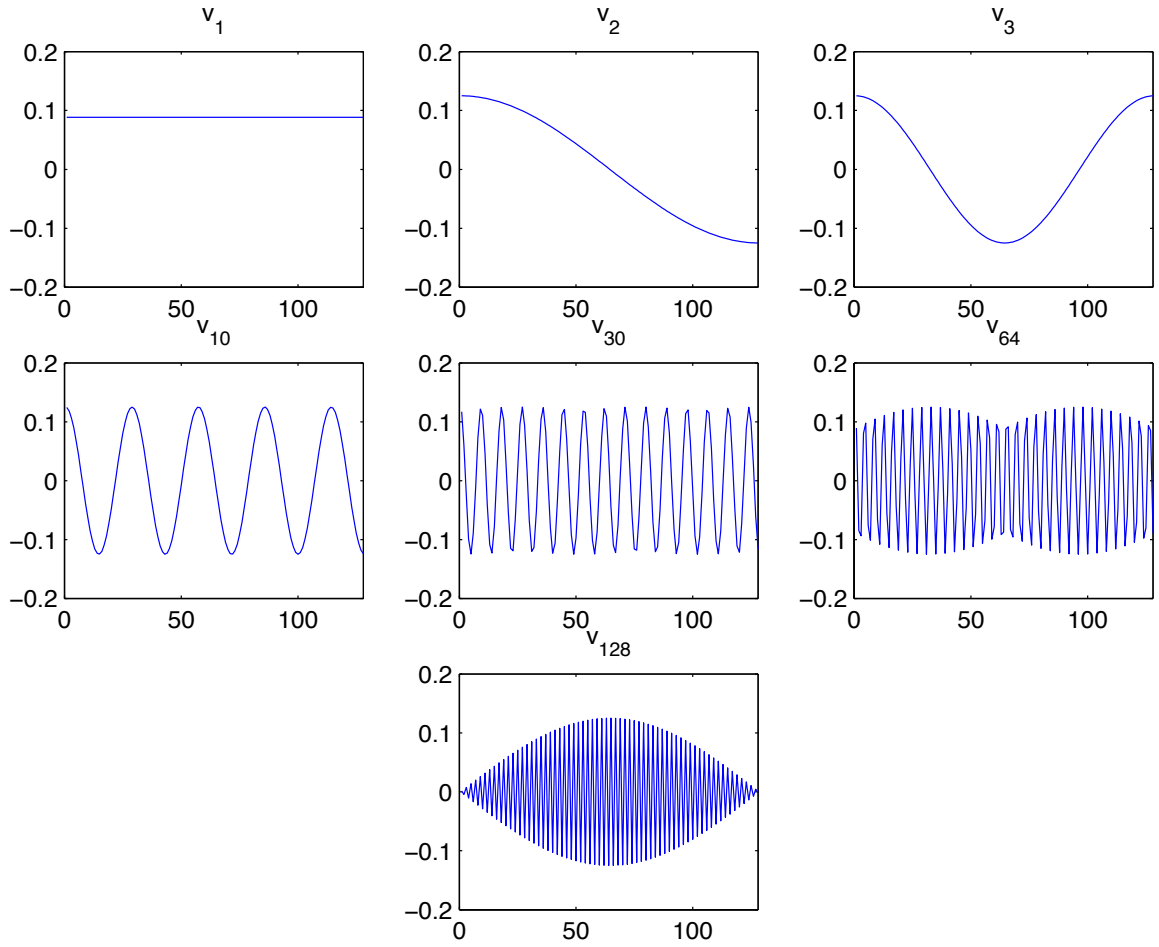


Figure 5.5: A few of the singular vectors for the matrices in Fig. 5.3

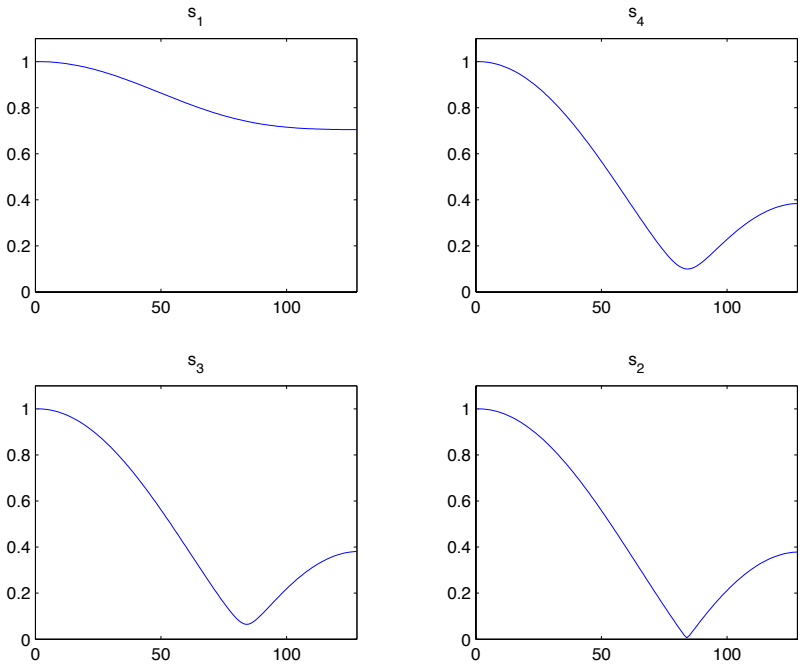


Figure 5.6: Singular value plots for the $\mathbf{K}$ matrices in Fig. 5.3

a zero singular value at around index 85. In signal-processing parlance, these linear operators look increasingly like low pass filters as we go from $i = 1$ to $i = 4$. While each matrix is, strictly speaking, invertible and while the four of them appear quite similar based on Figs. 5.3 and 5.4 from the structure of the singular values will have a rather substantial impact on the reconstruction results.

Here we look at the problem of recovering the $\mathbf{f}$ whose components are plotted in Fig. 5.7 both from noise free and noisy data. The clean data sets $\mathbf{g}_i = \mathbf{K}_i\mathbf{f}$ are shown in Fig. 5.8 while their noisy counterparts are displayed in Fig. 5.9. In each case the *same* noise vector is added to the clean data. The noise vector itself is comprised of zero mean, independent identically distributed Gaussian random variables with standard deviations equal to 0.04. As noted in the previous paragraph, $\mathbf{K}_i$ is technically invertible for each i so that solving the inverse problem here amounts to applying $\mathbf{K}_i^{-1}$ to each of the eight possible data vectors (four noiseless and four with additive noise). The results are shown in Figs. 5.10 and 5.11. In the case of the noise free data, the invertibility of $\mathbf{K}_i$ produces the anticipated results: perfect recovery of $\mathbf{f}$. The addition of noise however leads to substantial changes in the reconstruction. As the singular values structure comes closer and closer to possessing a zero, the influence of the noise on $\hat{\mathbf{f}}$ grows in a manner out of proportion to the size of the noise itself. More specifically, large amplitude, high frequency artifacts become increasingly dominant in the estimates of the object. This type of noise amplification is in fact seen across a broad range of problems, not just this somewhat artificial example, and is the primary characteristic of the ill-posed nature of these inverse problems.

To gain a more precise and more general understanding of the problem, let us continue to

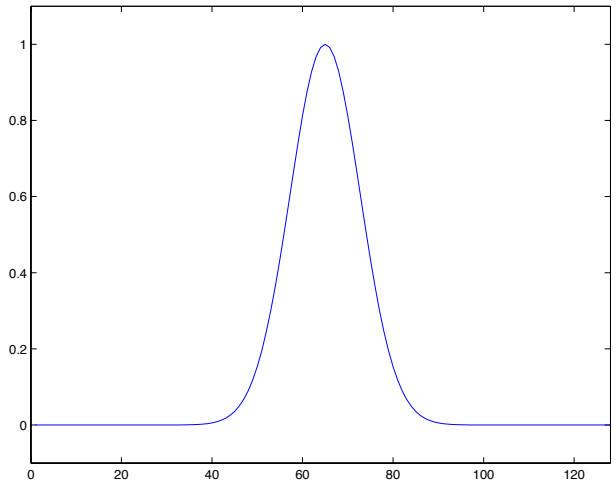


Figure 5.7: Object to be recovered

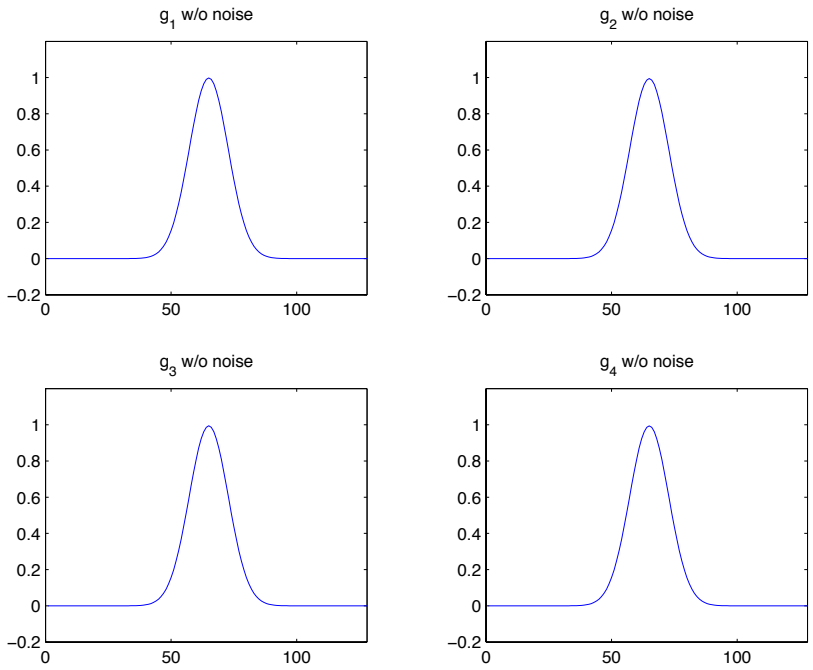


Figure 5.8: Noise free data

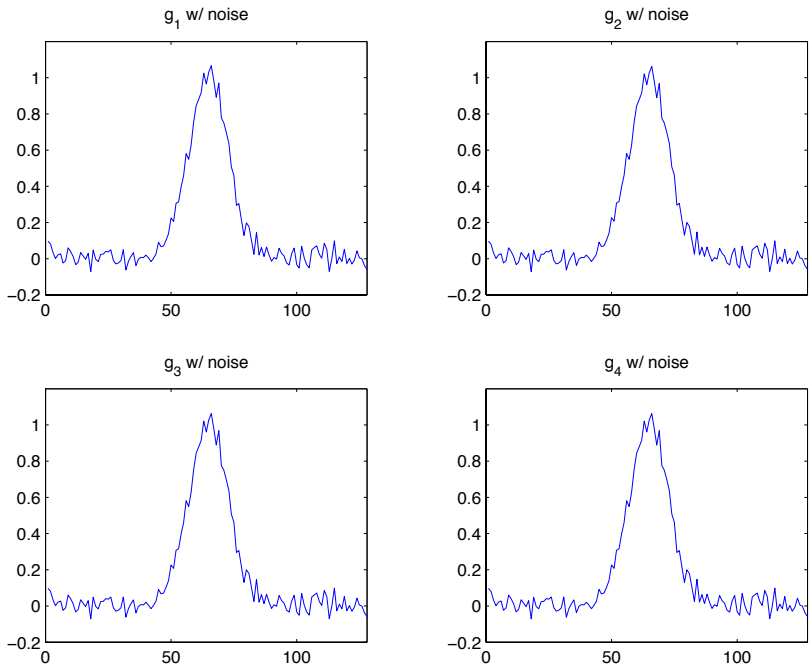


Figure 5.9: Noisy data

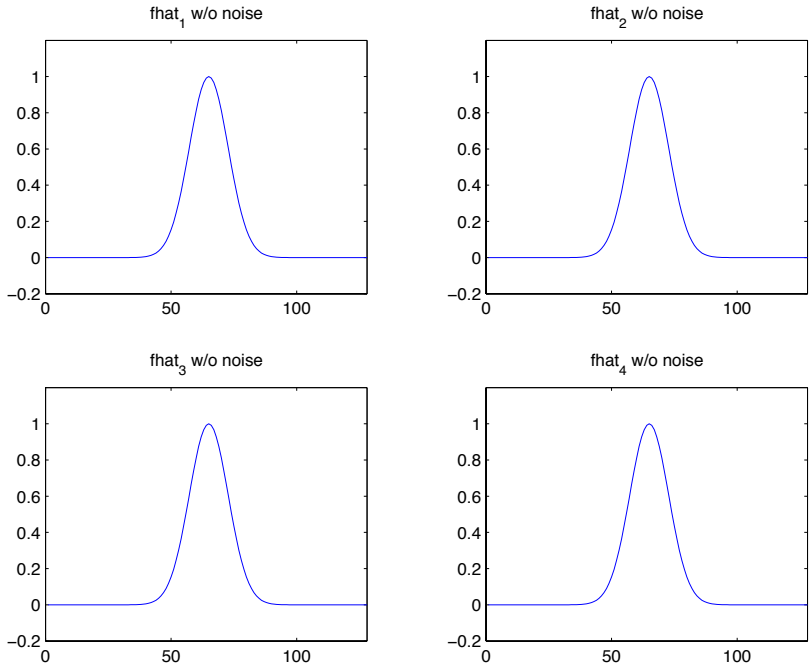


Figure 5.10: Noise free reconstructions

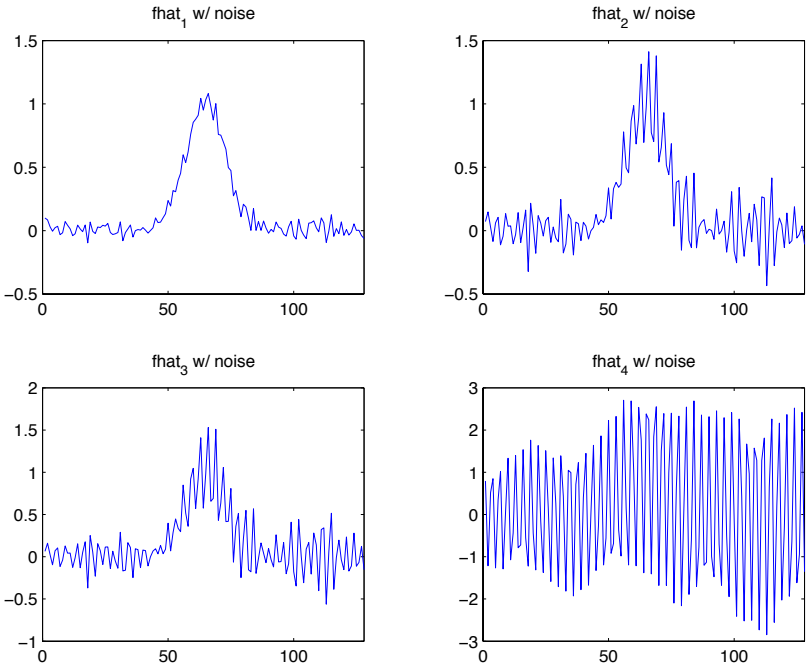


Figure 5.11: Noisy reconstructions

assume we have a square linear system, $M = N$, possessing a full set of N nonzero singular values, $\sigma_1 > \sigma_2 > \cdots > \sigma_N > 0$, but not necessarily having $\mathbf{U} = \mathbf{V}$. Exploiting the orthonormality of the $N \times N$ matrices $\mathbf{U}$ and $\mathbf{V}$, the inverse of $\mathbf{K} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ is just $\mathbf{K}^{-1} = \mathbf{V}\mathbf{\Sigma}^{-1}\mathbf{U}^T$ where $\mathbf{\Sigma}^{-1} = \text{diag}(\sigma_1^{-1}, \sigma_2^{-1}, \ldots, \sigma_M^{-1})$. Using this decomposition of the inverse of $\mathbf{K}$ means that

$$\hat{\mathbf{f}} = \mathbf{V}\mathbf{\Sigma}^{-1}\mathbf{U}\mathbf{g} = \sum_{k=1}^{N} \frac{\gamma_k}{\sigma_k}\mathbf{v}_k \tag{5.7}$$

where $\gamma_k = \mathbf{u}_k^T\mathbf{g}$ and $\mathbf{u}_k$ and $\mathbf{v}_k$ are the k-th columns of $\mathbf{U}$ and $\mathbf{V}$ respectively. We interpret (5.7) in terms of the analysis-filtering-synthesis interpretation of a linear operator discussed in § 2.2.3. That is, the reconstruction, $\hat{\mathbf{f}}$ is formed by the synthesis of a set of "modes," $\mathbf{v}_k$ each of which is weighted by a scaled generalized Fourier coefficient. The coefficients are the projection of $\mathbf{g}$ onto the basis given my the columns of $\mathbf{U}$ and the scaling is done using the inverse of the singular values of $\mathbf{K}$.

This interpretation implies that the action of the inverse of $\mathbf{K}$ is comprised of two "well-posed" steps, namely analysis and synthesis, between which is sandwiched the source of many of the difficulties for linear inverse problems, filtering. The orthonormality of $\mathbf{U}$ and $\mathbf{V}$ imply that their norms are unity. Thus there will not be any sensitivity or ill-posedness associated with these operations because they do not amplify (or for that matter attenuate) the "size" (as measured in the two-norm sense) of the vectors on which they act. Indeed, consider the case where the data vector $\mathbf{g}$ is perturbed by a small amount $\delta\mathbf{g}$ in that $\|\delta\mathbf{g}\| \ll 1$. By the triangle and Cauchy-Schwartz

inequalities, $\|\mathbf{U}(\mathbf{g}+\delta\mathbf{g})\| \leq \|\mathbf{U}\mathbf{g}\| + \|\mathbf{U}\delta\mathbf{g}\| \leq \|\mathbf{U}\|\|\mathbf{g}\| + \|\mathbf{U}\|\|\delta\mathbf{g}\|$. But $\|\mathbf{U}\| = 1$. Hence if $\|\delta\mathbf{g}\|$ is small, so too will be the change in the norm of the output.

The same is most definitely not true of the filtering operation. As indicated by (5.7), filtering is performed via multiplication by the diagonal matrix $\mathbf{\Sigma}^{-1}$ or equivalently, scaling each of the generalized Fourier coefficients γ_k by σ_k^{-1}. To see the impact of this scaling let us suppose that the data vector is $\mathbf{g} = \mathbf{K}\mathbf{f} + \mathbf{n}$ where each element of the noise vector $\mathbf{n}$ is independent of all the rest and distributed as a zero mean Gaussian random variable with variance ν^2. Letting $\hat{\mathbf{f}} = \mathbf{K}^{-1}\mathbf{g}$ and making use of (5.7) yields $\hat{\mathbf{f}} = \mathbf{f} + \mathbf{V}\mathbf{\Sigma}^{-1}\mathbf{U}^T\mathbf{n}$ so that the reconstruction error is given by $\mathbf{e} = \hat{\mathbf{f}} - \mathbf{f} = \mathbf{V}\mathbf{\Sigma}^{-1}\mathbf{U}^T\mathbf{n}$. To gauge the size of this error we examine the "average" value of $\|\mathbf{e}\|^2$. Using tools from standard statistical signal analysis [58, Section 3.5][3], this is easily shown to be

$$E\left\{\|\hat{\mathbf{f}} - \mathbf{f}\|_2^2\right\} = \sum_{k=1}^{N} \frac{\nu^2}{\sigma_k^2} \tag{5.8}$$

where $E\{\cdot\}$ is the expectation operator. In other words, even if ν is very small, on average the noise will not go to zero. In fact, the noise will generally be present for all k in the summation (5.7). Hence as the singular values go to zero, their inverse will be heading toward infinity thereby amplifying the impact of the corresponding $\mathbf{v}_k$ on $\hat{\mathbf{f}}$.

A commonly used tool for gauging the severity of the ill-posedness of $\mathbf{K}$ is the *condition number* of the matrix, $\kappa(\mathbf{K})$, defined as

$$\kappa(\mathbf{K}) = \|\mathbf{K}\|\|\mathbf{K}^{-1}\| \tag{5.9}$$

which can be show[4] to be equal to

$$\kappa(\mathbf{K}) = \frac{\sigma_1(\mathbf{K})}{\sigma_N(\mathbf{K})} = \frac{\text{Largest singular value of } \mathbf{K}}{\text{Smallest singular value of } \mathbf{K}}. \tag{5.10}$$

The utility of κ as a measure of ill-posedness arises from the following observation. Suppose that we have a linear system $\mathbf{K}\mathbf{f} = \mathbf{g}$ and we perturb $\mathbf{f}$ by $\delta\mathbf{f}$ so that $\mathbf{K}(\mathbf{f}+\delta\mathbf{f}) = \mathbf{g} + \delta\mathbf{g}$. From the triangle and Cauchy-Schwartz inequalities it is easily seen that

$$\|\delta\mathbf{f}\| \leq \|\mathbf{K}^{-1}\|\|\delta\mathbf{g}\|.$$

But since $\mathbf{K}\mathbf{f} = \mathbf{g}$, we see that $\|\mathbf{g}\| \leq \|\mathbf{K}\|\|\mathbf{f}\|$ or

$$\frac{1}{\|\mathbf{f}\|} \leq \|\mathbf{K}\|\frac{1}{\|\mathbf{g}\|}.$$

Hence

$$\frac{\|\delta\mathbf{f}\|}{\|\mathbf{f}\|} \leq \|\mathbf{K}^{-1}\|\frac{\|\delta\mathbf{g}\|}{\|\delta\mathbf{f}\|} \leq \|\mathbf{K}^{-1}\|\|\mathbf{K}\|\frac{\|\delta\mathbf{g}\|}{\|\delta\mathbf{g}\|} = \kappa(\mathbf{K})\frac{\|\delta\mathbf{g}\|}{\|\mathbf{g}\|}. \tag{5.11}$$

Thus, if we think of $\delta\mathbf{g}$ as the noise in the data, (5.11) indicates that the fractional change in the resulting $\mathbf{f}$ is bounded above by the relative power of the noise amplified by the condition number

[3] Will need an appendix on this stuff

[4] EXERCISE

of the matrix. In other words, the condition number places a bound on how perturbations in the data can be magniufied by the inversion process. As κ grows larger, the problem becomes more ill-posed and the impact of noise more pronounced. For the matrices in Fig. 5.3, the condition numbers are 1.41, 10.03, 15.14 and 160.86. In practice, condition numbers on the order of 10^15 are not uncommon. Hence even for a fairly mild condityion number of around 160, the impact of ill-posedness as seen in Fig. 5.11 can be substantial.

The issue of ill-posedness is quite closely related to those of existence and uniqueness. Keeping with the assumption that $N = M$, let us assume for a moment that rather than decaying to zero, the singular values were in fact equal to zero for all $i > i^*$. In this case $\mathbf{\Sigma}$ is of the form

$$\mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_1 & \mathbf{0}_{k^* \times N-k^*} \\ \mathbf{0}_{N-k^* \times k^*} & \mathbf{0}_{N-k^* \times N-k^*} \end{bmatrix} \tag{5.12}$$

where $\mathbf{0}_{m \times n}$ is the $m \times n$ matrix of all zeros and $\mathbf{\Sigma}_1 = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_{k^*})$. Drawing on the insight provided by (5.3) and (5.6), we see that (5.12) has elements of both non-uniqueness and non-existence. The bottom block row of zeros implies that in the presence of noise there will generally not be an $\mathbf{f}$ such that $\mathbf{g} = \mathbf{K}\mathbf{f}$. The block of zeros in the upper right block of $\mathbf{\Sigma}$ implies that if we satisfy ourself with ignoring the part of the problem associated with the bottom block of zeros, then the remaining problem is underdetermined, possesses a nullspace, and thus will not have a unique solution.

More generally one finds that the singular values decay toward zero, but are never exactly equal to zero. In those cases where there is a clear dividing line between "large and "small" values of σ_k, one could specify an effective i^* thereby reducing the problem to one where

$$\mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_1 & \mathbf{0}_{k^* \times N-i^*} \\ \mathbf{0}_{N-i^* \times i^*} & \mathbf{\Sigma}_{small} \end{bmatrix}. \tag{5.13}$$

Such a system possessed basically the same interpretation as (5.12) if one is willing to ignore $\mathbf{\Sigma}_{small}$. In most cases however, it is the unfortunate fact that no such clear division exists. Rather the decay of the singular values is gradual but unrelenting. Such problem do technically admit solutions to the extent that $\hat{\mathbf{f}} = \mathbf{K}^{-1}\mathbf{g}$ exists. However as we have seen, such solutions will be characterized by noise induced artifacts caused by the amplification of modes for which σ_k^{-1} are small, but not easily deemed negligible. Damping out these modes without totally ignoring their contribution is the goal regularization to be discussed in §XXX and XXX.

Before closing out this section, a couple of remarks are in order:

1. Eq. (5.8) indicates that the modes for which σ_k are significantly larger than ν will be impacted minimally by the noise. As we shall see throughout the remainder of this chapter, it is frequently the case that these modes contain generally low spatial frequency information. As the mode number i increases, it is often the case that the frequency content of the corresponding mode vector $\mathbf{v}_k$ also rises. Hence building on the case of diffraction tomography, one may conclude a bit more generally that for linear inverse problems the data provide stable and reliable information mostly about the low frequency content of the object.

2. If there were no noise and the data were exactly equal to $\mathbf{K}\mathbf{f}$ then the γ_k would also be going to zero "fast enough" to allow for a perfect reconstruction. If the noise is not of the white

Gaussian noise variety discussed above then the problem of ill-posedness *may* not arise. More specifically, if $\|\mathbf{u}_k^T\mathbf{g}\| \to 0$ at a rate faster than σ_k then the data are said to satisfy the *discrete Picard condition* [41] and the noise-induced amplification of some of the $\mathbf{v}_k$ will not occur.

5.2 The Pseudo-Inverse

An initial tool used to address the three issues of existence, uniqueness and (to a limited extent) stability is known as the *pseudo-inverse* of the matrix $\mathbf{K}$. In a discrete setting, the pseudo-inverse is equal to $\mathbf{K}^{-1}$ when this matrix exists. Thus the difficulties encountered with small but still nonzero singular values are not addressed using this inversion scheme and will be taken up in greater detail in § 5.3. When no inverse exists, the singular value matrix, $\mathbf{\Sigma}$, has the structure of (5.3) for full-rank overdetermined problems, (5.6) for full-rank underdetermined problems, or a variant of (5.12) for problems where $M \neq N$ and some of the singular values are zero. While we derive the pseudo-inverse in each of these cases separately in this section, the results are all remarkably similar and yield an inversion scheme which echoes (5.7):

$$\hat{\mathbf{f}} = \sum_{k\in\mathcal{K}} \frac{\mathbf{u}_k^T\mathbf{g}}{\sigma_k}\mathbf{v}_k = \mathbf{V}_1\mathbf{\Sigma}_1^{-1}\mathbf{U}_1^T\mathbf{g} \equiv \mathbf{K}^\dagger\mathbf{g} \tag{5.14}$$

where the index set $\mathcal{K} = \{k|\sigma_k \neq 0\}$ and $\mathbf{U}_1$ and $\mathbf{V}_1$ are comprised of the corresponding columns from the singular vector matrices $\mathbf{U}$ and $\mathbf{V}$. In other words, the pseudo-inverse constructs an estimate $\hat{\mathbf{f}}$ which is quite similar as that which is obtained when $\mathbf{K}$ is invertible, just restricted to the subspaces of $\mathbf{U}$ and $\mathbf{V}$ for which the singular values are non-zero.

5.2.1 Full-Rank Overdetermined Inverse Problems

As illustrated in Fig. 5.1, problems in this class are characterized by an inability to find any $\mathbf{f}$ such that $\mathbf{Kf}$ is equal to the data vector $\mathbf{g}$ because the columns of the $M \times N$ matrix $\mathbf{K}$ with $M > N$ span a subspace of $\mathbb{R}^N$. The pseudo-inverse of $\mathbf{K}$ for this problem is obtained by seeking that $\mathbf{f}$ such that $\mathbf{Kf}$ is as close to $\mathbf{g}$ as is possible. Formally, we have

$$\hat{\mathbf{f}} = \arg\min_{\mathbf{f}} \|\mathbf{g} - \mathbf{Kf}\|_2^2 \tag{5.15}$$

Since $\mathbb{R}^N$ is a Hilbert space, the projection theorem guarantees that a unique solution to (5.15) will exist and moreover, provide a means of finding that solution. Specifically, $\hat{\mathbf{f}}$ will be that $\mathbf{f}$ which makes the error $\mathbf{g} - \mathbf{Kf}$ orthogonal to any element in the range of $\mathbf{K}$. This means that for any $\boldsymbol{\phi} \in \mathbb{R}^N$ we must have $(\mathbf{g} - \mathbf{Kf}) \perp \mathbf{K}\boldsymbol{\phi}$. This is accomplished if and only if

$$(\mathbf{K}\boldsymbol{\phi})^T(\mathbf{g} - \mathbf{K}\hat{\mathbf{f}}) = \boldsymbol{\phi}^T(\mathbf{K}^T\mathbf{g} - \mathbf{K}^T\mathbf{K}\hat{\mathbf{f}}) = 0. \tag{5.16}$$

Because $\boldsymbol{\phi}$ is arbitrary, this condition is satisfied if and only if

$$\mathbf{K}^T\mathbf{K}\hat{\mathbf{f}} = \mathbf{K}^T\mathbf{g} \to \hat{\mathbf{f}} = (\mathbf{K}^T\mathbf{K})^{-1}\mathbf{K}^T\mathbf{g} \tag{5.17}$$

the linear system to the right of the arrow in (5.17) is known as the *normal equations* and the $\hat{\mathbf{f}}$ solving this system is known as the *linear least squares* solution to the overdetermined problem. Finally, the matrix taking the data to $\hat{\mathbf{f}}$, $\mathbf{K}^\dagger = (\mathbf{K}^T\mathbf{K})^{-1}\mathbf{K}^T$ is the pseudo-inverse of $\mathbf{K}$ for the full rank overdetermined system.

Another interpretation of this solution is obtained by examining Fig. 2.4. By writing the data space, $\mathbb{R}^M$ as the direct sum of two orthogonal subspaces $\mathcal{R}(\mathbf{K})$ and $\mathcal{R}^\perp(\mathbf{K}) = \mathcal{N}(\mathbf{K}^T)$ the condition of making $\mathbf{g} - \mathbf{K}\mathbf{f}$ as small as possible amounts to requiring that there be no component of this vector in the range of $\mathbf{K}$. Were such a component present then we must be able to adjust how we combine the vectors spanning this space in order to remove this error. The requirement that no part of $\mathbf{g} - \mathbf{K}\mathbf{f}$ be in $\mathcal{R}(\mathbf{K})$ means that *all* of the error must lie in $\mathcal{R}^\perp(\mathbf{K}) = \mathcal{N}(\mathbf{K}^T)$. From a linear algebraic perspective, this means $\mathbf{K}^T(\mathbf{g} - \mathbf{K}\mathbf{f}) = 0$, as seen above.

Still a third approach to this solution follows from the second. Specifically, we decompose $\mathbf{g}$ uniquely as $\mathbf{g}_r + \mathbf{g}_0$ where $\mathbf{g}_r \in \mathcal{R}(\mathbf{K})$ and $\mathbf{g}_0 \in \mathcal{R}^\perp(\mathbf{K})$. Since $\mathbf{g}_r$ is in the range of $\mathbf{K}$, there should be a unique solution to $\mathbf{K}\mathbf{f} = \mathbf{g}_r$. In fact this will be the case. To see this we start by recalling that $\mathbf{g}_r = \mathbf{P}_{\mathcal{R}(\mathbf{K})}\mathbf{g}$ where $\mathbf{P}_{\mathcal{R}(\mathbf{K})}$ is the orthogonal projector onto the range of $\mathbf{K}$. From the discussion on page 28, an orthonormal basis for $\mathcal{R}(\mathbf{K})$ is given by $\mathbf{U}_1$, the set of left singular vectors associated with the nonzero singlar values of $\mathbf{K}$. According to Example 2.19, given such a basis for $\mathcal{R}(\mathbf{K})$, the projector is constructed as $\mathbf{P}_{\mathcal{R}(\mathbf{K})} = \mathbf{U}_1\mathbf{U}_1^T$. So, making use of the SVD of $\mathbf{K}$, the problem we wish to solve is:

$$\mathbf{U}\begin{bmatrix}\mathbf{\Sigma}_1\\ \mathbf{0}\end{bmatrix}\mathbf{V}^T\hat{\mathbf{f}} = \mathbf{U}_1\mathbf{U}_1^T\mathbf{g} \tag{5.18}$$

but writing $\mathbf{U}$ as $[\mathbf{U}_1\ \mathbf{U}_2]$, and taking advantage of the orthonormality of this matrix we have

$$\mathbf{U}^{-1}\mathbf{U}_1 = \mathbf{U}^T\mathbf{U}_1 = \begin{bmatrix}\mathbf{U}_1^T\\ \mathbf{U}_2^T\end{bmatrix}\mathbf{U}_1 = \begin{bmatrix}\mathbf{I}\\ \mathbf{0}\end{bmatrix} \tag{5.19}$$

Using (5.19) as well as the fact that $\mathbf{\Sigma}^\dagger = \begin{bmatrix}\mathbf{\Sigma}_1^{-1}\ \mathbf{0}\end{bmatrix}$ is a left inverse of $\mathbf{\Sigma}$ for this problem, (5.18) can be solved as

$$\begin{aligned}\hat{\mathbf{f}} &= \mathbf{V}\begin{bmatrix}\mathbf{\Sigma}_1\ \mathbf{0}\end{bmatrix}\mathbf{U}^T\mathbf{U}_1\mathbf{U}_1^T\mathbf{g}\\ &= \mathbf{V}\begin{bmatrix}\mathbf{\Sigma}_1\ \mathbf{0}\end{bmatrix}\begin{bmatrix}\mathbf{I}\\ \mathbf{0}\end{bmatrix}\mathbf{U}_1^T\mathbf{g} = \mathbf{V}\mathbf{\Sigma}_1^{-1}\mathbf{U}_1^T\mathbf{g}\end{aligned} \tag{5.20}$$

thereby showing how to find $\mathbf{V}_1$, $\mathbf{U}_1$ and $\mathcal{K}$ in (5.14). Additionally since

$$\mathbf{\Sigma}_1^{-1}\mathbf{U}_1^T = \begin{bmatrix}\mathbf{\Sigma}_1^{-1}\ \mathbf{0}\end{bmatrix}\begin{bmatrix}\mathbf{U}_1^T\\ \mathbf{U}_2^T\end{bmatrix} = \mathbf{\Sigma}^\dagger\mathbf{U}^T$$

we have

$$\hat{\mathbf{f}} = \mathbf{V}\mathbf{\Sigma}^\dagger\mathbf{U}^T\mathbf{g} \tag{5.21}$$

which provides another expression for $\mathbf{K}^\dagger$ in terms of the components of the SVD of $\mathbf{K}$.

5.2.2 Full-Rank Underdetermined Inverse Problems

The development of the pseudo-inverse for the full-rank underdetermined is quite complementary to that of the overdetermined case. As discussed on page 88, the primary issue here is the presence of the nullspace for the matrix $\mathbf{K}$ so that many $\mathbf{f}$'s exist such that $\mathbf{Kf} = \mathbf{g}$. As in the previous section, the pseudo-inverse is constructed by defining $\hat{\mathbf{f}}$ as the solution to an optimization problems. Here thought the problem is to select the "smallest" $\mathbf{f}$ which is still consistent with the data as in

$$\begin{aligned} \hat{\mathbf{f}} &= \arg\min_{\mathbf{f}} \|\mathbf{f}\|_2^2 \\ &\text{subject to } \mathbf{Kf} = \mathbf{g}. \end{aligned} \tag{5.22}$$

The solution to (5.22) is called the *minimum-norm* (or just *min-norm*) solution to the underdetermined, full-rank problem $\mathbf{Kf} = \mathbf{g}$.

Much as we solved the overdetermined problem by decomposing the data space, $\mathbb{R}^M$ into two orthogonal pieces, here we do the same but for the object space, $\mathbb{R}^N$. That is we write any $\mathbf{f} \in \mathbb{R}^N$ that solves $\mathbf{Kf} = \mathbf{g}$ as the unique sum of two components: $\mathbf{f}_n \in \mathcal{N}(\mathbf{K})$ and a second piece $\mathbf{f}_r \in \mathcal{N}^\perp(\mathbf{K}) = \mathcal{R}(\mathbf{K}^T)$. That $\mathbf{f}$ with no component in the nullspace of $\mathbf{K}$ is the solution to (5.22)[5]. Because the remaining part of the solution $\mathbf{f}_r$ lies in the range of $\mathbf{K}^T$, we have $\mathbf{f}_r = \mathbf{K}^T\mathbf{x}$ for some vector $\mathbf{x} \in \mathbb{R}^M$. Thus, $\mathbf{g} = \mathbf{KK}^T\mathbf{x}$ so $\mathbf{x} = (\mathbf{KK}^T)^{-1}\mathbf{g}$ and finally

$$\hat{\mathbf{f}} = \mathbf{K}^T(\mathbf{KK}^T)^{-1}\mathbf{g} \tag{5.23}$$

is the solution to (5.22). Assuming $\mathbf{K}$ has full row rank, the inverse in (5.23) must exist and the solution is unique[6]. Also, from (5.23)this we see that the pseudo-inverse of $\mathbf{K}$ for the underdetermined full rank problem is

$$\mathbf{K}^\dagger = \mathbf{K}^T(\mathbf{KK}^T)^{-1}. \tag{5.24}$$

To see the role played by the SVD in this solution, we start by noting that the SVD of $\mathbf{K}$ for this problem takes the form

$$\mathbf{K} = \mathbf{U}\,[\mathbf{\Sigma}_1\ \mathbf{0}] \begin{bmatrix} \mathbf{V}_1^T \\ \mathbf{V}_2^T \end{bmatrix}$$

and that the columns of $\mathbf{V}_1$ form an orthonormal basis for $\mathcal{N}^\perp(\mathbf{K})$. So $\mathbf{f}_r = \mathbf{V}_1\mathbf{x}$ and we have

$$\mathbf{U}\,[\mathbf{\Sigma}_1\ \mathbf{0}] \begin{bmatrix} \mathbf{V}_1^T \\ \mathbf{V}_2^T \end{bmatrix} \mathbf{V}_1\mathbf{x} = \mathbf{g}. \tag{5.25}$$

From (5.25) we conclude that $\mathbf{x} = \mathbf{\Sigma}_1^{-1}\mathbf{U}^T\mathbf{g}$ so

$$\begin{aligned} \hat{\mathbf{f}} &= \mathbf{V}_1\mathbf{\Sigma}_1^{-1}\mathbf{U}^T\mathbf{g} = [\mathbf{V}_1\ \mathbf{V}_2] \begin{bmatrix} \mathbf{\Sigma}_1^{-1} \\ \mathbf{0} \end{bmatrix} \mathbf{U}^T\mathbf{g} \\ &\equiv \mathbf{V}\mathbf{\Sigma}^\dagger\mathbf{U}^T\mathbf{g}. \end{aligned} \tag{5.26}$$

Hence again, $\mathbf{K}^\dagger = \mathbf{V}\mathbf{\Sigma}^\dagger\mathbf{U}^T$ only now the pseudo-inverse of $\mathbf{\Sigma}$ is appropriately altered to take into account the underdetermined structure of the problem.

[5]EXERCISE: Prove this using Pythagoras?

[6]EXERCISE: Prove this.

5.2.3 Reduced Rank Problems

In the most general case where there are $P < \min(M, N)$ nonzero singular values, $\mathbf{\Sigma}$ is of the form

$$\mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_1 & \mathbf{0}_{12} \\ \mathbf{0}_{21} & \mathbf{0}_{22} \end{bmatrix} \tag{5.27}$$

with $\mathbf{\Sigma} = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_P)$ and the sizes of the zero block are dependent on the value of P and whether $M < N$, $M > N$, or $M = N$.[7] The SVD of $\mathbf{K}$ is now written as

$$\mathbf{K} = [\mathbf{U}_1 \; \mathbf{U}_2] \begin{bmatrix} \mathbf{\Sigma}_1 & \mathbf{0}_{12} \\ \mathbf{0}_{21} & \mathbf{0}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{V}_1^T \\ \mathbf{V}_2^T \end{bmatrix}. \tag{5.28}$$

Following the discussion on page 97, the bottom block row of $\mathbf{\Sigma}$ indicates that in general $\mathbf{g}$ will not be in the range of $\mathbf{K}$. The upper right block of zeros shows that even if $\mathbf{g}$ were in (or were made to be in) $\mathcal{R}(\mathbf{K})$, there would still be a nullspace to the problem so non-uniqueness would be an issue. Because this class of problems is a blend of that seen in the previous two subsections, it should come as no surprise that the pseudo-inverse is derived using elements of both previous cases.

To be more precise, $\hat{\mathbf{f}}$ here is obtained as the min-norm solution to the linear problem where the data $\mathbf{g}$ is projected into the range of $\mathbf{K}$. Formally we have

$$\begin{aligned} \hat{\mathbf{f}} &= \arg\min_{\mathbf{f}} \|\mathbf{f}\|_2^2 \\ &\text{subject to } \mathbf{K}\mathbf{f} = \mathbf{P}_{\mathcal{R}(\mathbf{K})}\mathbf{g}. \end{aligned} \tag{5.29}$$

Using the methods from § 5.2.2and § 5.2.3, the unique solution to (5.29) is obtained using (5.28) along with the following two facts

- $\mathbf{P}_{\mathcal{R}(\mathbf{K})} = \mathbf{U}_1\mathbf{U}_1^T$
- $\hat{\mathbf{f}}$ must be of the form $\mathbf{V}_1\mathbf{x}$.

The resulting linear system for $\mathbf{x}$ is

$$[\mathbf{U}_1 \; \mathbf{U}_2] \begin{bmatrix} \mathbf{\Sigma}_1 & \mathbf{0}_{12} \\ \mathbf{0}_{21} & \mathbf{0}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{V}_1^T \\ \mathbf{V}_2^T \end{bmatrix} \mathbf{V}_1\mathbf{x} = \mathbf{U}_1\mathbf{U}_1^T\mathbf{g}$$

from which we obtain $\mathbf{x} = \mathbf{\Sigma}_1^{-1}\mathbf{U}_1^T\mathbf{g}$ so

$$\hat{\mathbf{f}} = \mathbf{V}_1\mathbf{\Sigma}_1^{-1}\mathbf{U}_1^T\mathbf{g} \tag{5.30}$$

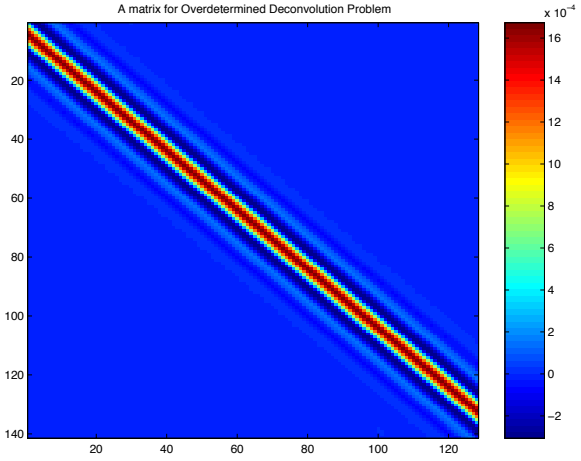
or equivalently

$$\hat{\mathbf{f}} = [\mathbf{V}_1 \; \mathbf{V}_2] \begin{bmatrix} \mathbf{\Sigma}_1^{-1} & \mathbf{0}_{21}^T \\ \mathbf{0}_{12}^T & \mathbf{0}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{U}_1^T \\ \mathbf{U}_2^T \end{bmatrix} \mathbf{g} \equiv \mathbf{V}\mathbf{\Sigma}^\dagger\mathbf{U}^T\mathbf{g} \tag{5.31}$$

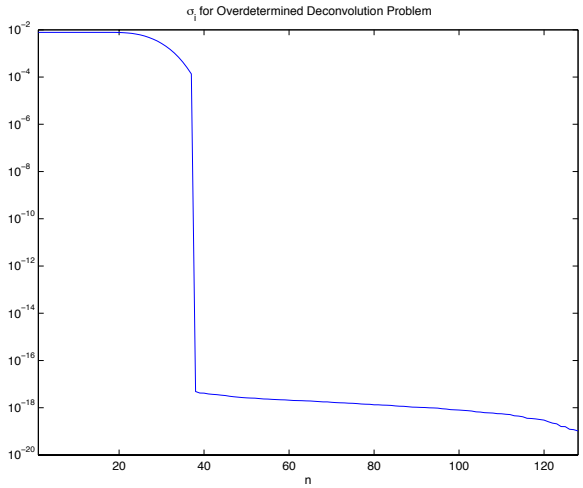
just as in (5.21)and (5.26).

To illustrate the performance of the pseudo-inverse and motivate the need for additional work in stabilizing the solution to linear inverse probelm, we consider a number of examples motivated by deconvolution and inversion of the Born approximation.

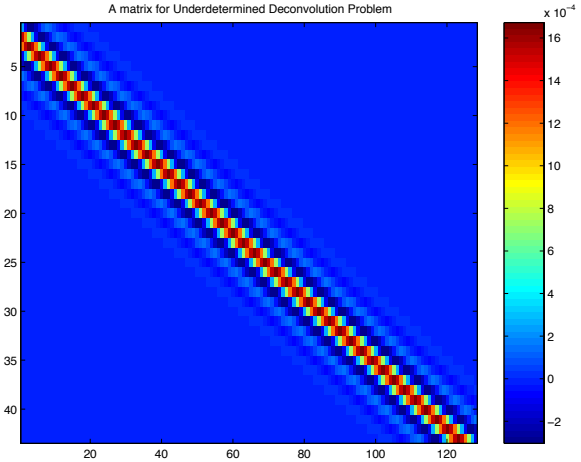
[7] EXERCISE: Work out the dimensions in all three cases.



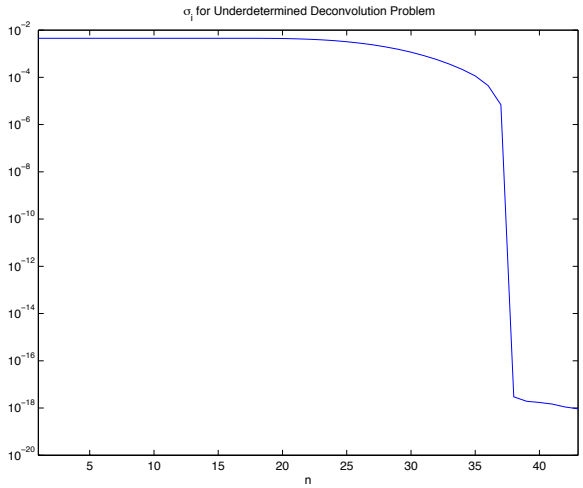
(a) **K** matrix for overdetermined deconvolution problem



(b) Singular values for overdetermined deconvolution problem



(c) **K** matrix for overdetermined deconvolution problem



(d) Singular values for underdetermined deconvolution problem

Figure 5.12: Matrices and singular value structures for over and underdetermined deconvolution type examples

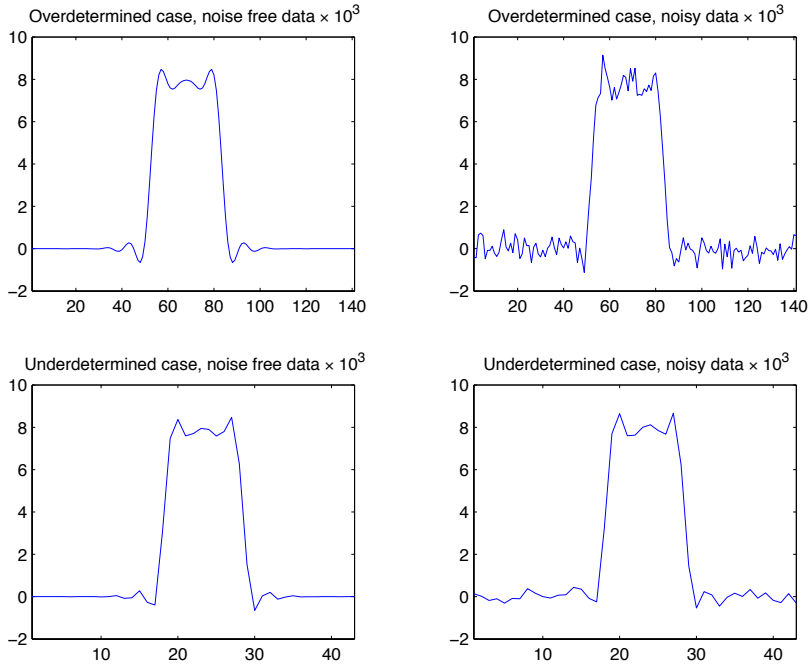


Figure 5.13: Data vectors for deconvolution type problem.

In Fig. 5.12(a) and (c) the $\mathbf{K}$ matrices for a pair of deconvolution-type problems are shown. The matrix in (a) is of size 141×128 resulting in a slightly over determined problem. The one shown in Fig. 5.12(b) is obtained by removing every third row from that in (a) and hence results in an underdetermined problem. The structures of the singular values for each of these matrices are shown in Fig. 5.12(b) and (d). Note the differences in the x axes for both of these plots which in turn reflect the differing sizes of the underlying $\mathbf{K}$ matrices. In both cases there are slightly fewer than 40 significant singular values. Despite the fact that the first matrix has more rows than columns, the singular value plot indicates that almost two thirds of the singular values are negligible resulting in what will turn out to be a rather substantial "numerical" nullspace.

The "Toeplitz" type of structure associated with the $\mathbf{K}$ matrices indicate that the singular vectors $\mathbf{u}_k$ and $\mathbf{v}_k$ will be more or less sinusoidal and of increasing frequency as i increases. Coupled with the singular value plots, we can conclude that these matrices act as low-pass filters. That this is true is evident from Fig. 5.13 where the data vectors are shown for these two problems. In the left row are the noise-free data. Data with a low level of additive Gaussian noise are shown on he right. In all cases the true $\mathbf{f}$ is shown in the top panel of Fig. 5.14. We see from Fig. 5.13(a) and (c) that the data look very much like a low-pass filtered version of the input "box" function. Specifically, recalling that the Fourier transform of a box in a sinc function, we know that the discontinuities in the box are manifest in the Fourier domain by a slow $1/\omega$ type of decay in the amplitude of the Fourier transform. Low pass filtering a box then would remove the high frequency information needed to build the edges resulting in ringing (or Gibbs phenomenon) in the output of the filter. This is precisely what we see in Fig. 5.13.

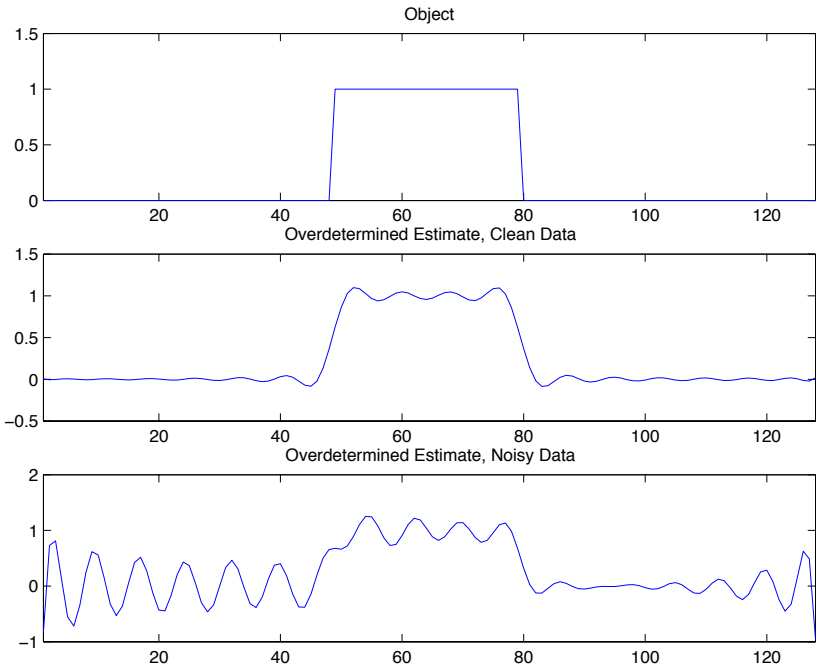


Figure 5.14: Inversion results for overdetermined deconvolution-type problem

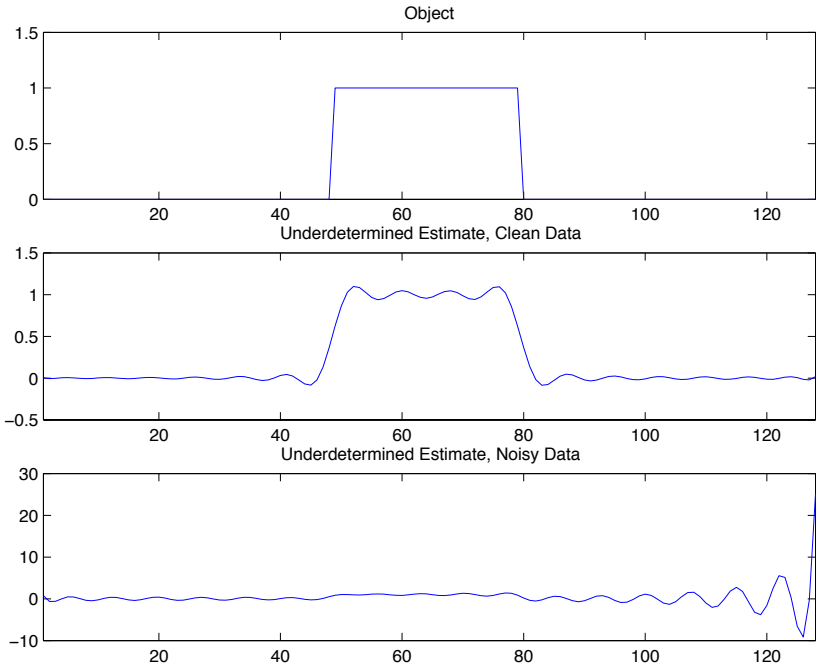


Figure 5.15: Inversion results for underdetermined deconvolution-type problem

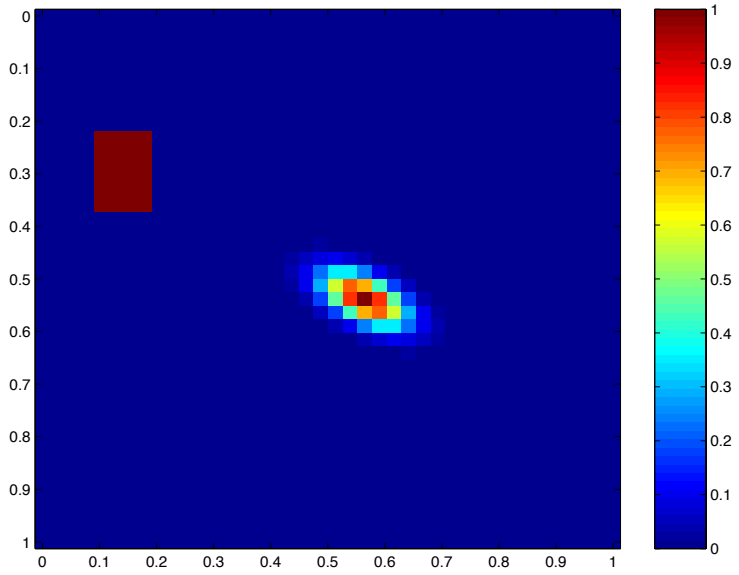


Figure 5.16: Object to be recovered in Born inversion example

The results of inverting the data in Fig. 5.13 via the psedo inverse are shown in Fig. 5.14 for the matrix in Fig. 5.12(a) and in Fig. 5.15 for the matrix in Fig. 5.12(c). In each plot the upper panel shows the object, the middle panel is the result of applying the pseudo-inverse to data with no noise and the bottom panels are the data with noise reconstructions. In computing the pseudo-inverse, we set to zero all singular values less than 10^{-16} in size. Hence in both cases, we are really plotting min-norm least squares reconstruction as described in § 5.2.3.

For noise-free data, the reconstructions in both cases are again low pass versions of the true box profile. This is entirely consistent with the mathematics of the pseudo-inverse which says that only non-zero singular values and their associated singular vectors participate in the reconstruction. For the specific problems shown here, the non-zero singular values were those associated with low-frequency (low-pass) singular vectors. Hence, following the same reasoning as was applied in understanding Fig. 5.13, we observe that the middle panels in Figs. 5.14 and 5.15 show reconstructions which look like low frequency approximations to the box.[8]

For the noisy data, the bottom panels in Figs. 5.14 and 5.15 indicate that the quality of the noise-free reconstructions is obtained specifically because there is no little high-frequency component in the data. The addition of white noise to the data increases the amount of signal in the higher frequency range from what the pseuso-inverse is "expecting" based on the structure of $\mathbf{f}$. While the increase is fairly small (compare the noisy and noise-free data in Fig. 5.13), its impact is quite severe especially in the underdetermined case. As we discussed in § 5.1, these large amplitude, high frequency artifacts are classic examples of the impact of ill-posedness (ill-conditioning) on the reconstructions.

We next consider the recovery of the two dimensional object shown in Fig. 5.16 using data

[8]Perhaps plot projection of the box onto the low frequency subspaces to make this point even clearer.

obtained in an scattering framework under the Born approximation. All units are normalized so that the region to be images is 1×1. Ten transmitters are arrayed along the left side of the medium and ten receivers along the right edge. Referring to Table 3.1, we assume an acoustics type of inverse problems so that the background $k_b = \omega/c_b$ and $k_s(\mathbf{r}) = \omega f(\mathbf{r})$ where $c_b = 1$ is the normalized background sound speed and $f(\mathbf{r})$ is the object to be recovered. Data sets are collected for $\omega \in \{5, 10, 15, 20\}$. The region to be imaged is decomposed into a 40×40 grid of pixels and (for simplicity, not accuracy) the basis functions needed to discretize the Born integral equation are taken to be δ functions located at the centers of each pixel.

The data for each frequency are plotted in Fig. 5.17. Because the Born matrix is complex valued, each observation provides two pieces of data for the inversion: the real part and the imaginary component. We consider the result of inverting with each of the data sets alone (for which we have $10 \times 10 \times 2 = 200$ pieces of information as well as with data from multiple frequencies. In the multi-frequency case, we form four data sets comprised of information from the following sets of frequencies: $\{5\}$, $\{5, 10\}$, $\{5, 10, 15\}$, and $\{5, 10, 15, 20\}$ yielding matrices with $200, 400, 600$ and 800 rows respectively.

The singular values plots of the resulting eight linear systems are shown in Fig. 5.18. Unlike the previous example where a clear distinction exists between singular values that may be considered significant from those that are clearly negligible, no such separation occurs here. In all eight cases, the singular values decay gradually from their maximum values to values close to machine precision. Also, all eight problems are underdetermined in that there are 1600 pixel values to be determined given at most 800 observations.

The pseudo-inverse results for this example are shown in Figures 5.19– 5.22. As with the deconvolution problem, the presence of noise significantly degrades the results (note the differences on the colorbar axes). In the noise-free case, there are no great differences in each of the single frequency results. In all cases, the edges of the square block are somewhat blurred while the shape of the smoother object in the middle is fairly well captured. The difficulty in capturing the edges of the block are a two dimensional example of the same phenomena seen in the deconvolution problem: an ability of the pseudo-inverse to recover accurately only low frequency structural information. Finally, we note the presence of ringing artifacts in the backgrounds of the higher frequency reconstructions.

The addition of multiple frequencies in Fig. 5.20 helps to significantly reduce the ringing and provides improved recovery of the edges of the block object on the left side of the region. Direct comparison of the single frequency and multi-frequency results though is a bit subtle. In the later case, the amount of data is greater and the singular value plots in Fig. 5.18 show that these additional data are significant in that they allow for the recovery of information about $\mathbf{f}$ using far more singular vectors. A fairer and perhaps more enlightening comparison would be one in which the size of the multi-frequency data sets was kept the same as each of the single frequency inversions. Also, it would be interesting to study the utility of adding sources and receivers that encircle a larger portion of the region to be imaged.[9]

[9]EXERCISES

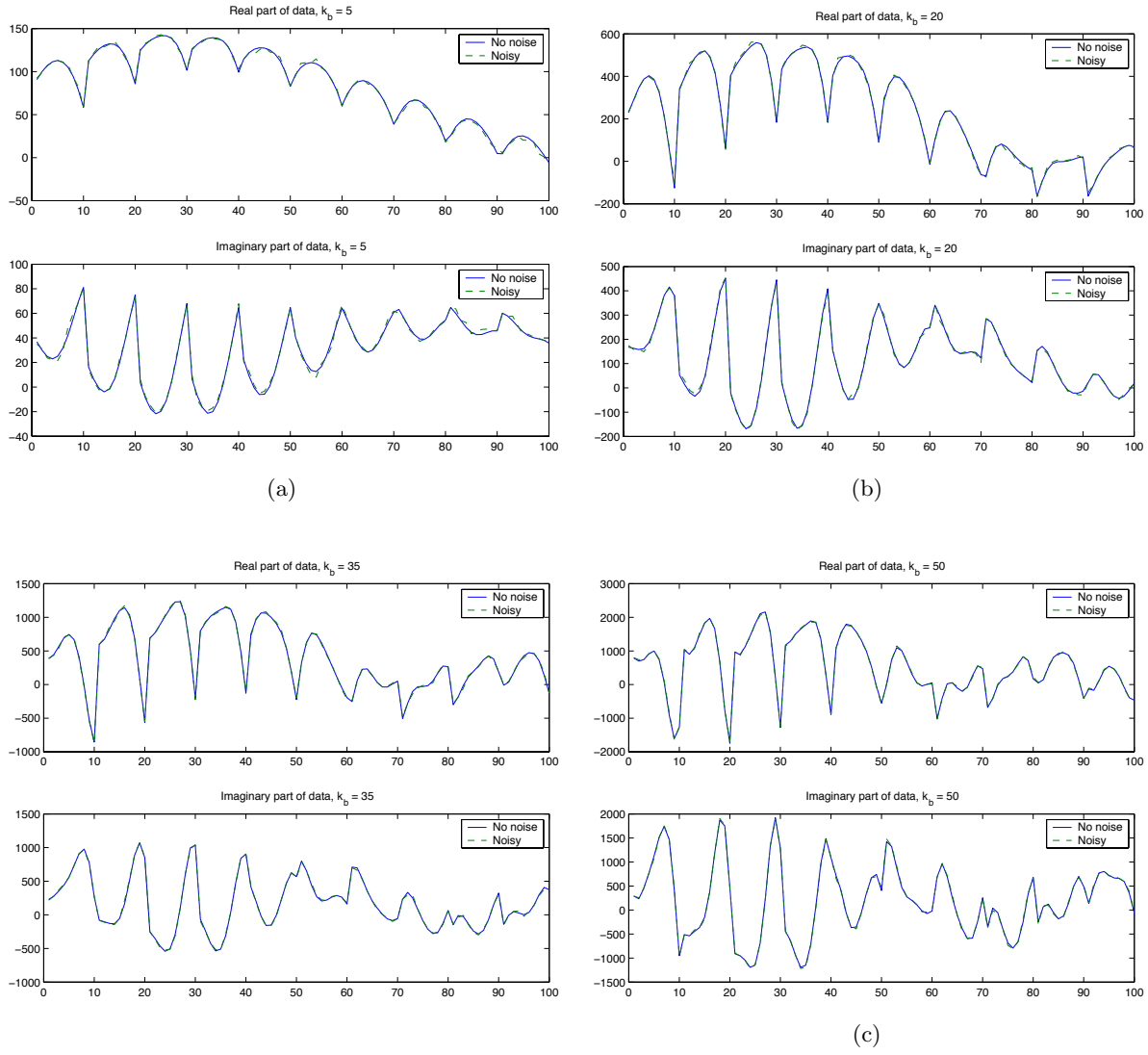
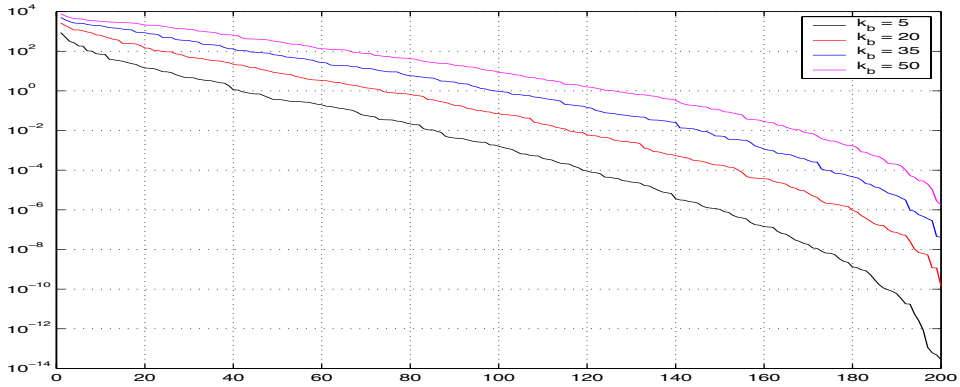
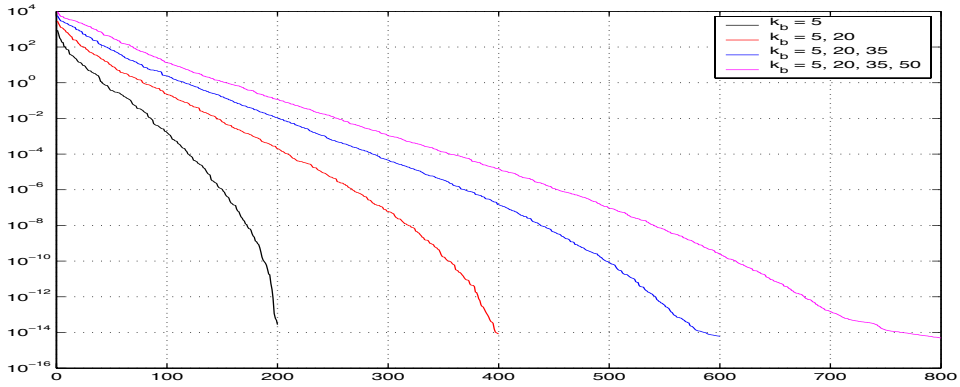


Figure 5.17: Data sets used for Born inversion



(a)



(b)

Figure 5.18: Singular value plots for the Born inverse problems

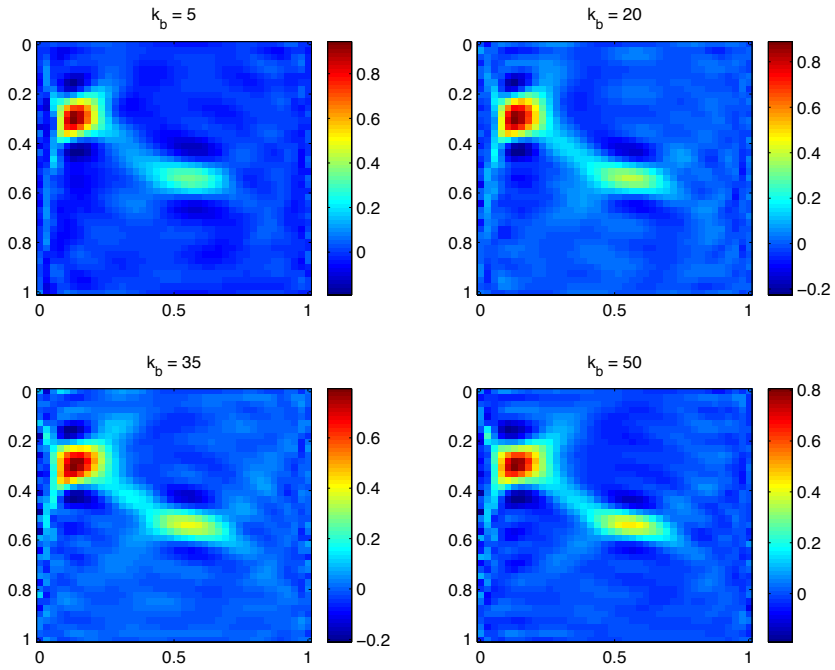


Figure 5.19: Inversion results with single frequency noise free data

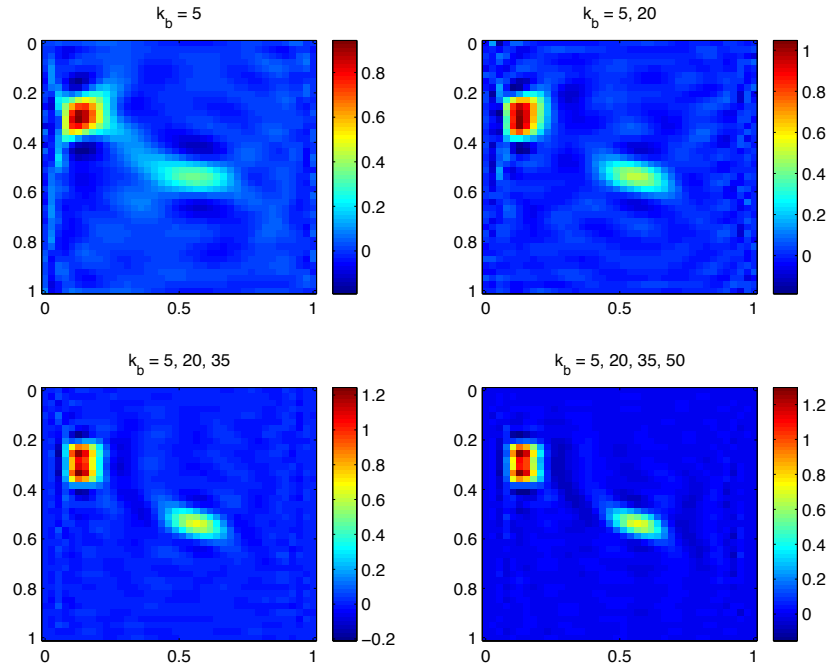


Figure 5.20: Inversion results with multi-frequency noise free data

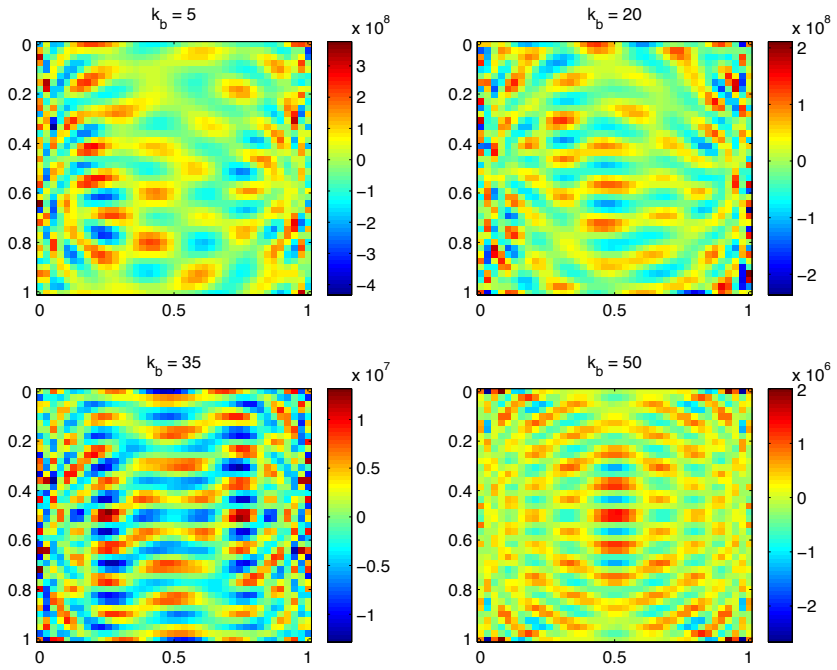


Figure 5.21: Inversion results with single frequency noisy data

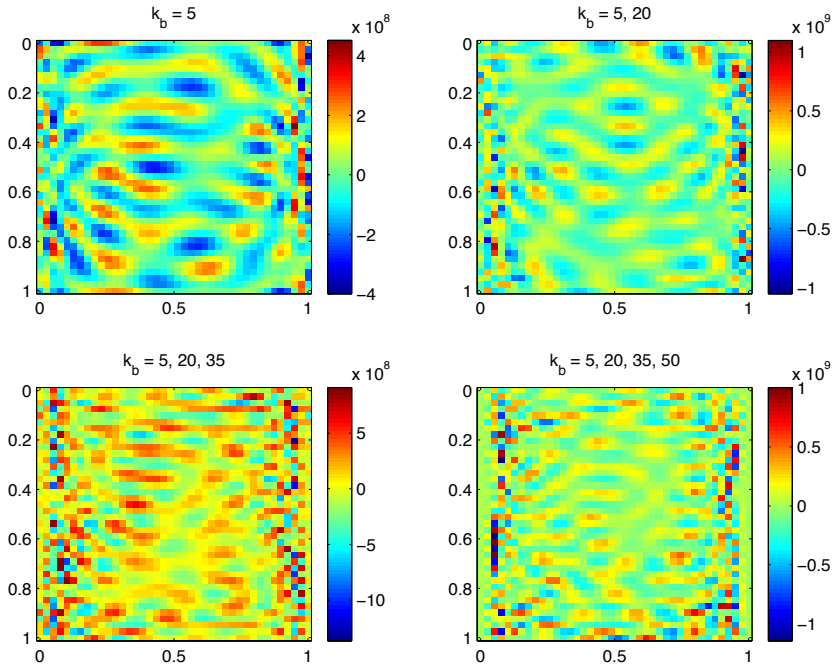


Figure 5.22: Inversion results with multi-frequency noisy data

5.3 Regularization I

Thus far we have encountered two types of inverse procedures: exact inverse methods such as filtered back-projection/propagation and the pseudo-inverse. While exact methods are appropriate for data rich problems, their performance can suffer for cases where data are limited. The pseudo-inverse provides an initial means of obtaining a reconstruction when either no exact solution exists or an infinite number of such solutions can be found. As indicated by the example discussed at the end of the last section, the pseudo-inverse works quite well for data limited problems where there is no noise in the observations; however the presence of noise can severely degrade the results obtained using this method.

In essence this lack of robustness to noise is directly related to the manner in which the pseudo-inverse treats the singular values of $\mathbf{K}$. As long as a singular value is not zero, it and its associated singular vectors plays a role in the reconstruction. Thus even small amounts of noise in modes associated with small singular values is significantly amplified. This then suggests one very obvious modification to the pseudo-inverse: eliminate the contribution of singular values whose size is less than some predetermined threshold. The resulting approach opens the door to a sequence of *regularization* approaches for adding more robustness to inversion than is available from the pseudo-inverse. An initial set of regularization schemes which fit into the framework of linear inverse problems are presented here. The topic of regularization is taken up again in § 6.2 in the case of nonlinear problems.

The notion of a regularization procedure can be made quite formal and leads to some very interesting and non-trivial analysis especially problems where neither object nor the data are functions are discrete [28]. Here we content ourselves with an introduction to this theory appropriate for cases where the data are discrete and the object may be either finite or infinite dimensional [5–7]. Thus let us take K to be a linear operator mapping a Hilbert space X into $\mathbb{R}^M$. By the structure of the problem, the ith datum, g_i is the inner product of the object f with some linear functional ϕ_i. For $f \in \mathbb{R}^N$, the functional can be thought of as the ith row of the matrix $\mathbf{K}$. When f is an element of an infinite dimensional Hilbert space such as L^2, following the inner product takes the form of $\int \phi_i(\mathbf{r}) f(\mathbf{r})\, d\mathbf{r}$.

A *regularization algorithm* is defined as a mapping R_λ from Y to X which is dependent on a *regularization parameter* $\lambda > 0$ satisfying three properties [5]:

1. For any $\lambda > 0$, $\mathcal{R}(R_\lambda)$ is contained within the span of the ϕ_i.
2. For $\lambda > 0$ $\|R_\lambda\| \leq \|K^\dagger\|$
3. The following limit holds :
$$\lim_{\lambda \downarrow 0} \|R_\lambda - K^\dagger\| = 0 \tag{5.32}$$

Intuitively, these conditions can be seen to make a good deal of sense in terms. Recall that for the discrete problem the span of the ϕ_i is captured by the $\mathbf{v}_i$. Hence, we see in (5.14) that $\hat{\mathbf{f}}$ is a linear combination of these vectors. So, the first condition indicates that a good regularization scheme is one where this property is retained. The second condition requires that the regularizer do no worse than the pseudo-inverse in terms of its noise amplification. Finally, the third states that as the regularization parameter decreases to zero, we should recover the pseudo-inverse.

This approach to defining a regularization scheme reserves a special place for the pseudo-inverse. Indeed, in the more functional theoretic approaches to regularization developed for example in [28], the pseudo-inverse plays something of the role of a gold standard. To understand this, we know that the need for regularization comes from the presence of noise in the data. Within the structure of that theory, were it not for this noise, the "best" that we could do in terms of inversion is to apply the pseudo-inverse to the data. A good regularization method then is one which converges to the pseudo-inverse in the limit of small noise and hence vanishing λ.

5.3.1 The Truncated Singular Value Decomposition

As we just indicated, perhaps the most obvious way of moving past the strict definition of the pseudo-inverse is to ignore the contribution from singular values which can safely be regarded as small. That is we define the reconstruction as

$$\hat{\mathbf{f}} = \sum_{k=1}^{k_0} \frac{1}{\sigma_k} \left(\mathbf{u}_k^T \mathbf{g}\right) \mathbf{v}_k. \tag{5.33}$$

In this case, k_0 is known as a *regularization parameter*. Its presence here represents a fairly major philosophical departure in inversion schemes from what we have so far been discussing. Exact inverse methods and the pseudo-inverse all derive their structures entirely from the physics of the sensing modality. With the introduction of k_0 here we now have a means of controlling the inverse procedure which is independent of the physics; that is, entirely user defined. While a number of somewhat rigorous techniques for algorithmically selecting regularization parameters are discussed in § 5.3.5, in many practical cases, some level of user intervention is really required to choose the "best" one.

The inversion scheme in (5.33) is known as the *truncated singular value decomposition* (TSVD). Although it was motivated by qualitative arguments, in fact there is a sense in which the TSVD is optimal. Specifically, it is know that the closest matrix $\mathbf{B}$ of rank k_0 approximating $\mathbf{K}$ in the induced two norm sense is

$$\mathbf{B}^* = \underset{\text{rank}(\mathbf{B})=k_0}{\arg\min} \|\mathbf{B} - \mathbf{K}\|_2^2 = \sum_{k=1}^{k_0} \sigma_k \mathbf{u}_k \mathbf{v}_k^T. \tag{5.34}$$

Thus, (5.33) can be viewed as the pseudo-inverse for the problem where $\mathbf{K}$ is replaced by its closest rank k_0 approximation.

The advantages of the TSVD are its ease of implementation as well as its strong performance for problems where the choice of k_0 is not difficult. Specifically, we would expect that the TSVD would perform best when a clear distinction can be made between significant and negligible singular values. Such is certainly the case for problem whose σ_k have a sharp cutoff as in Fig. 5.12. For problems whose singular values decay gradually, as in Fig. 5.18, no obvious threshold exists. Hence, choosing k_0 becomes a more delicate, perhaps subjective, exercise. Such problems highlight the primary shortcoming of the TSVD: much like the pseudo-inverse, the contributions of the subspaces of $\mathbf{U}$ and $\mathbf{V}$ associated with the truncated singular values are completely absent from $\hat{\mathbf{f}}$. Thus, any important information contained in these subspaces is also lost. To address this issue requires a

procedure in which these subspaces are allowed to play a limited role in the structure of $\hat{\mathbf{f}}$. Their impact must be controled however to avoid the noise amplification issue seen with the pseudo-inverse.

5.3.2 Spectral Filtering

A first step in this process is again a somewhat natural approach to extending the pseudo-inverse and indeed the TSVD. We can view the TSVD as a windowing of the singular values of $\mathbf{K}$ where the window is either 1 for $1 \leq k \leq k_0$ and 0 for $k > k_0$. To moderate the impact of this sharp cutoff, we can construct inversion schemes using more general window functions:

$$\hat{\mathbf{f}} = \sum_{k=1}^{\min(M,N)} w_{\lambda,k} \frac{1}{\sigma_k} \left(\mathbf{u}_k^T \mathbf{g}\right) \mathbf{v}_k \tag{5.35}$$

where $w_{\lambda,k}$ defines the weight given to each singular value and λ is a regularization parameter governing the shape of the window. Thus, by appropriately designing $w_{\lambda,k}$, we now have a mechanism for including in a controlled manner information from all components of $\mathbf{U}$ and $\mathbf{V}$. To ensure that (5.35) fits within the definition of a regularization procedure given on page 111, we require two conditions on the values $w_{\lambda,k}$

1. For any $\lambda > 0$, and all i $0 \leq w_{\lambda,k} \leq 1$
2. For all k
$$\lim_{\lambda \to} w_{\lambda,k} = 1$$

The first condition says that the weights serve to attenuate as opposed to amplify the impact of the small singular values thereby ensuring that the conditioning is improved relative to the pseudo-inverse. The second implies that as the regularization parameter goes to zero, we should recover the pseudo-inverse as is required by the third condition on page 111.

A few examples of valid window function include:

Example 5.1 The flat top function:

$$w_{\lambda,k} = \begin{cases} 1 & i \leq \lceil \frac{1}{\lambda} \rceil \\ 0 & i > \lceil \frac{1}{\lambda} \rceil \end{cases} \tag{5.36}$$

where $\lceil x \rceil$ is the first integer larger than x. This window is identical to the TSVD where $k_0 = \lceil \frac{1}{\lambda} \rceil$.

Example 5.2 The triangle window is formally defined as

$$w_{\lambda,k} = \begin{cases} 1 - \frac{k-1}{\lceil \frac{1}{\lambda} \rceil} & i \leq \lceil \frac{1}{\lambda} \rceil \\ 0 & i > \lceil \frac{1}{\lambda} \rceil \end{cases} \tag{5.37}$$

Shown in Fig. 5.23 are two examples of triangular windows; one for $\lambda = 1/6$ and the other with $\lambda = 1/20$. As λ goes to zero it is not hard to show that $w_{\lambda,k} \to 1$.

Example 5.3 As shown in Fig. 5.24 a decaying exponential[10] can quite easily be used as a window

[10] EXERCISE: Use tanh as a window to approximate the flat top

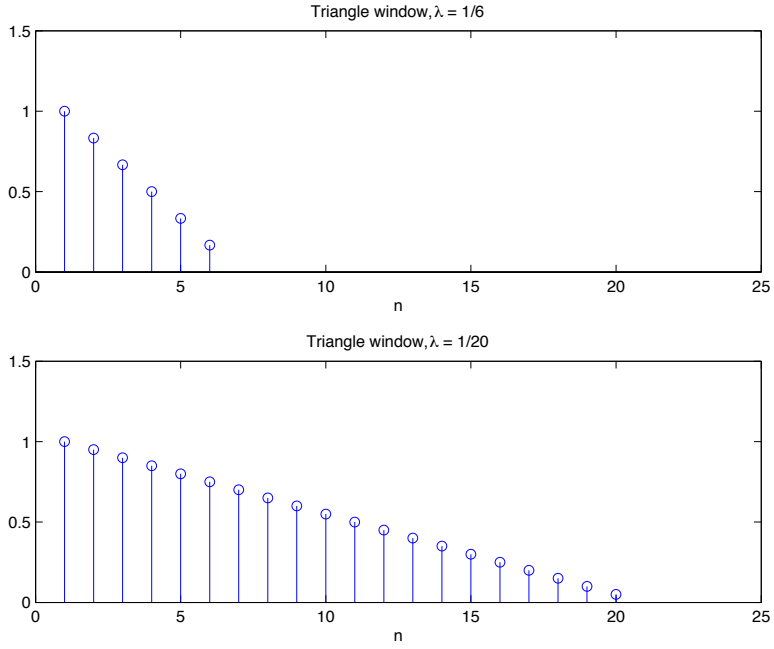


Figure 5.23: Triangle windows

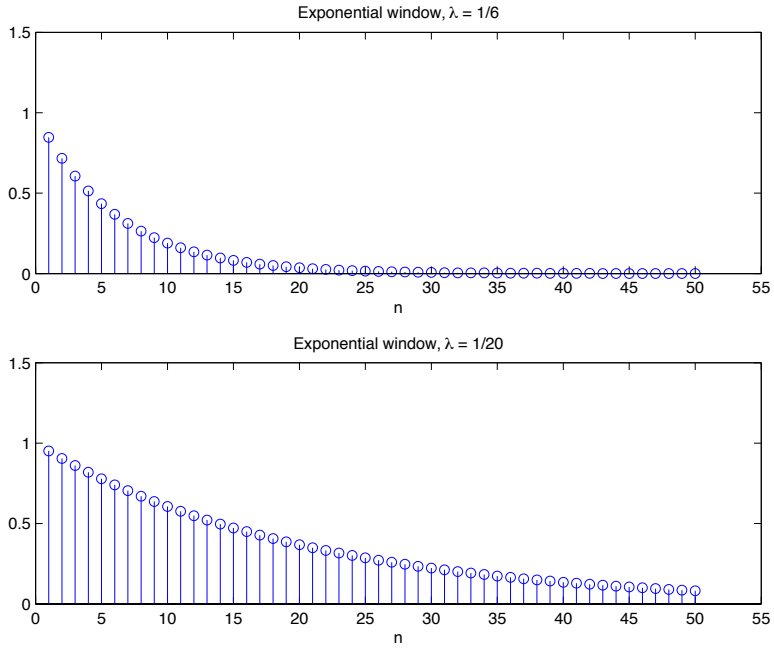


Figure 5.24: Exponential windows

function. Here we simply take

$$w_{\lambda,k} = \exp(k\lambda) \tag{5.38}$$

Example 5.4 Perhaps the most important window function we will encounter is fundamentally different from the above three in that its structure is dependent on the singular values in the following manner

$$w_{\lambda,k} = \frac{\sigma_k^2}{\sigma_k^2 + \lambda}. \tag{5.39}$$

It is easily seen that (5.39) defines a valid window function. Clearly, For $\lambda > 0$, $w_{\lambda,k} \in [0, 1]$ and $w_{\lambda,k} = 1$ for $\lambda = 0$. The impact on the reconstruction is most clearly seen by substituting (5.39) into (5.35) to arrive at

$$\hat{\mathbf{f}} = \sum_k \frac{\sigma_k}{\sigma_k^2 + \lambda} \left(\mathbf{u}_k^T \mathbf{g}\right) \mathbf{v}_k. \tag{5.40}$$

From (5.40) we see that as when $\sigma_k \to 0$, the weight provided that mode in the reconstruction also goes to zero as it should. By the same token, when σ_k is large relative to the regularization parameter, λ,

$$\frac{\sigma_k}{\sigma_k^2 + \lambda} \to \frac{1}{\sigma_k}$$

indicating that such contributions to $\hat{\mathbf{f}}$ are treated the in the same manner here as in the pseudo-inverse.

5.3.3 Variational Regularization Methods

The TSVD and spectral filtering methods both approach the problem of improved robustness to noise in terms of modifications to the matrix $\mathbf{K}$; specifically, its singular value structure. An alternate idea (and one which we shall see is more closely related to windowing than one might first think) comes from thinking about the problem in terms of properties of $\mathbf{f}$. Looking at the inversion results in the bottom panels of Figs. 5.14 and 5.15 as well as those if Figs. 5.21 and 5.22, we see that the artifacts produced by the pseudo-inverse come in the form of large amplitude, high frequency corruption in $\hat{\mathbf{f}}$. In most applications however we have prior expectations or even hard constraints on the behavior of the unknown object. In the case of X-ray tomography for example, the $\mathbf{f}$ is the density of the material being scanned and hence cannot assume negative values. In geophysics problems and certain classes of nondestructive evaluation problems, one model for the subsurface is a collection of layers with more or less constant properties (sound speed, electrical conductivity etc.). In cases where properties do vary spatially in addition to non-negativity, one would expect gradual variations in some range of bounded amplitudes. In other words, even if we did not know the true distributions of $\mathbf{f}$ the results in Figs. 5.14, 5.15, 5.21 and 5.22 would still be rejected for being not natural, not in line with our prior expectations concerning how $\mathbf{f}$ should behave.

The issue we face then is how to incorporate this prior information into the inversion process in some quantitative way. To do this, we extend the variational approach used to define the pseudo-inverse in (5.22) and (5.29) to include analytically defined constraints that capture our prior knowledge. In the abstract, there are a number of ways this can be accomplished. It may

be possible to define a class of functions, $\mathcal{C}$, whose behavior reflects our expectiations. A suitable inverse then is one belonging to this class *and* providing a good fit to the data:

$$\hat{\mathbf{f}} = \arg\min_{\mathbf{f}\in\mathcal{C}} \|\mathbf{g} - \mathbf{K}\mathbf{f}\|_2^2. \tag{5.41}$$

While it is not hard, at least formally, to define classes for *e.g.* positive functions, bandlimited functions, etc. [84], solving the resulting optimization problem is not at all straightforward.

Here and into the next chapter, we look at function classes whose mathematical structure is based on a norm of some function of $\mathbf{f}$ as a measure of the size of an undesirable feature of the object. For example one way of quantifying the notion that the amplitude of $\mathbf{f}$ should not be large is to say that $\|\mathbf{f}\|_2^2 = \mathbf{f}^T\mathbf{f}$ should be less than some value. Similarly, the highly oscillatory artifacts are manifest in the size not of $\mathbf{f}$, but of its derivative, or gradient in multiple dimensions. To make this clearer recall that the derivative of $\sin(\omega x)$ is $\omega\cos(\omega x)$. So the higher the frequency ω, the larger is the derivative. Constraining these oscillations amounts to a restriction on the $\|\mathbf{L}\mathbf{f}\|_2^2$ where $\mathbf{L}$ is matrix approximation to the gradient[11].

The use of the two norm in the above paragraph is neither necessary nor, in some important cases, useful as we shall see in Chapter 6. So to be a bit more general, let $\rho(\mathbf{f})$ be a norm-based measure of the size of $\mathbf{f}$. There are three variationally-based ways we can think of defining $\hat{\mathbf{f}}$ using both ρ and the information present in the data.

Approach 1	**Approach 2**	**Approach 3**
$\hat{\mathbf{f}} = \arg\min_{\mathbf{f}} \|\mathbf{g} - \mathbf{K}\mathbf{f}\|_2^2$	$\hat{\mathbf{f}} = \arg\min_{\mathbf{f}} \rho(\mathbf{f})$	$\hat{\mathbf{f}} = \arg\min \|\mathbf{g} - \mathbf{K}\mathbf{f}\|_2^2 + \lambda\rho(\mathbf{f})$
subject to $\rho(\mathbf{f}) \leq \lambda$	subject to $\|\mathbf{g} - \mathbf{K}\mathbf{f}\|_2^2 \leq \lambda$	

Comparing Approaches 1 and 2, we see that one can think of the data error term $\|\mathbf{g} - \mathbf{K}\mathbf{f}\|$ as a type of constraint on the same footing as ρ. This is captured explicitly in Approach 3 in which $\hat{\mathbf{f}}$ is defined as a solution to a minimization problem comprised of two terms; one encouraging that $\hat{\mathbf{f}}$ be faithful to the data while second requiring that the reconstruction have few artifacts as captured by the smallness of ρ.

As discussed at some depth in [5, Section V-A] the first two approaches have the same solution which in fact is obtained for cases where the "$\leq$" constraint is satisfied with equality[12]. Now, the solution of an equality constrained optimization problem is typically found though the use of Lagrange multiplier methods. Such problems possess the same functional form as that in Approach 3 with the primary difference being that the Lagrange multiplier, λ is also determined to ensure that $\rho(\mathbf{f}) = \lambda$ or $\|\mathbf{g} - \mathbf{K}\mathbf{f}\| = \lambda$ depending on the problem. In essence then, the basic variational problem we wish to solve for $\hat{\mathbf{f}}$ given some λ is that of Approach 3. When it some time to choose the value for λ, we can do so to enforce one of the constraints in Approaches 1 or 2. Alternatively, abandoning the interpretation of λ as a Lagrange multiplier and rather thinking of it as a parameter used to balance the value of ρ against that of $\|\mathbf{g} - \mathbf{K}\mathbf{f}\|$, we can use one of the regularization parameter selection methods to be discussed in § 5.3.5.

[11] EXERCISE: Build gradient matrix in one and two dimensions

[12] This follows from fact that the norm function is convex

5.3.4 Tikhonov-type Methods

Here we consider problems where $\rho(\mathbf{f})$ can be written as a quadratic function of the elements of $\mathbf{f}$, that is $\rho(\mathbf{f}) = \mathbf{f}^T\mathbf{Q}\mathbf{f}$ for some matrix $\mathbf{Q}$. This choice is motivated both by the fact that such constraints end up being both reasonable and useful and because the resulting variation problem has a closed form solution. Per the discussion in the last section, for the vast majority of cases of any practical use, the matrix $\mathbf{Q}$ arises from a discretization of one or more differential operators designed to penalize large amplitude and high frequency oscillations in $\hat{\mathbf{f}}$; that is

$$\rho(\mathbf{f}) = \|\mathbf{L}\mathbf{f}\|_2^2 = \mathbf{f}^T\mathbf{L}^T\mathbf{L}\mathbf{f} \tag{5.42}$$

from which we can identify $\mathbf{Q} = \mathbf{L}^T\mathbf{L}$. Choosing $\mathbf{L} = \mathbf{I}$ yield what is known as *Tikhonov regularization*. In addition to the identity, another common choice is to take $\mathbf{L}$ to be a discretized form of a gradient opertator. In a discrete setting, the gradient does not strictly exist; however the idea of a derivative as a measure of the local change in a sequence of numbers can be formaalized. The simplest such measure in one dimension at least is $\mathbf{f}_{k+1} - \mathbf{f}_k$ for $k = 1, 2, \ldots, N-1$, the difference between adjacent elements of the vector $\mathbf{f}$ which gives rise to the $N-1 \times N$ matrix $\mathbf{L}$:

$$\mathbf{L} = \begin{bmatrix} -1 & 1 & 0 & \ldots & 0 & 0 \\ 0 & -1 & 1 & \ldots & 0 & 0 \\ & & \ddots & \ddots & & \\ 0 & 0 & 0 & \ldots & -1 & 1 \end{bmatrix} \tag{5.43}$$

Equation (5.43) is by no mans the only choice we have for a derivative-like regularization matrix. For example, recalling the discussion in § 3.4, we could use (3.32) as the basis for constructing a $\mathbf{L}$ that implements a centered difference approach to the derivative. Similarly, constructing high fidelity filters for approximating in a discrete sense a derivative is a well known problem in the field of digital signal processing [77, Section 8.2.5] and thus could play a role here as well.[13]

It is possible that one would want a regularizer that guards against say r classes of artifacts in $\mathbf{f}$. For example, both the two norm of $\mathbf{f}$ as well as its derivative may need to be controlled. In general, assuming we can find one $\mathbf{L}_i$ for each then by defining

$$\mathbf{L} = \begin{bmatrix} \alpha_1\mathbf{L}_1 \\ \alpha_2\mathbf{L}_2 \\ \vdots \\ \alpha_r\mathbf{L}_r \end{bmatrix} \tag{5.44}$$

we still fall within the mathematical structure of (5.42) because

$$\begin{aligned} \|\mathbf{L}\mathbf{f}\|_2^2 &= \mathbf{f}^T\mathbf{L}^T\mathbf{L}\mathbf{f} \\ &= \mathbf{f}^T\left(\sum_{k=1}^{r}\alpha_k^2\mathbf{L}_k^T\mathbf{L}_k\right)\mathbf{f} \\ &= \sum_{k=1}^{r}\alpha_k^2\mathbf{f}^T\mathbf{L}_k^T\mathbf{L}_k\mathbf{f} = \sum_{k=1}^{r}\|\alpha_k\mathbf{L}_k\mathbf{f}\|_2^2 \equiv \rho(\mathbf{f}) \end{aligned} \tag{5.45}$$

[13] EXERCISE: Other options for derivative operators

where the α_k may be used to weight the different $\mathbf{L}_k$ as needed or desired. Thus, the basic mathematical form of the problem we wish to solve is

$$\hat{\mathbf{f}} = \arg\min_{\mathbf{f}} \|\mathbf{g} - \mathbf{K}\mathbf{f}\|_2^2 + \lambda\|\mathbf{L}\mathbf{f}\|_2^2 \tag{5.46}$$

with the understanding that $\mathbf{L}$ may well represent a concatenation of more basic regularizers as in (5.44). Finally, we shall refer to (5.46) as a Tikhonov regularized solution to the inverse problems with the recognition that, strictly speaking, this term applies only to the case where $\mathbf{L} = \mathbf{I}$ and we should be using a more cumbersome phase such as generalized Tikhonov inversion.

There are two methods we shall present for solving (5.46). First, by the same argument as we used in (5.45) to argue that the stacked $\mathbf{L}_i$ had the same mathematical structure vis a vis the two norm as a single $\mathbf{L}$, we can write (5.46) as

$$\hat{\mathbf{f}} = \arg\min_{\mathbf{f}} \|\tilde{\mathbf{g}} - \tilde{\mathbf{K}}\mathbf{f}\|_2^2 \tag{5.47}$$

where

$$\tilde{\mathbf{g}} = \begin{bmatrix} \mathbf{g} \\ 0 \end{bmatrix} \qquad \tilde{\mathbf{K}} = \begin{bmatrix} \mathbf{K} \\ \sqrt{\lambda}\mathbf{L} \end{bmatrix} \tag{5.48}$$

Eq. (5.47) though is of the same form as the linear least squares problem in (5.15) where the data vector and system matrix have been augmented to include the regularizer. But we know the solution to be

$$\hat{\mathbf{f}} = \left(\tilde{\mathbf{K}}^T\tilde{\mathbf{K}}\right)^{-1}\tilde{\mathbf{K}}^T\tilde{\mathbf{g}} = \left(\mathbf{K}^T\mathbf{K} + \lambda\mathbf{L}^T\mathbf{L}\right)^{-1}\mathbf{K}^T\mathbf{g} \tag{5.49}$$

We can also arrive at (5.49) using more traditional optimization methods. According to basic multivariate calculus, the extrema of an objective function can be found by solving the equations that result when the gradient of the function with respect to the unknowns is set equal to zero. In the case of (5.46), we start by expanding the objective function as

$$\begin{aligned} C(\mathbf{f}) &\equiv (\mathbf{g} - \mathbf{K}\mathbf{f})^T(\mathbf{g} - \mathbf{K}\mathbf{f}) + \lambda^2\mathbf{L}^T\mathbf{L} \\ &= \mathbf{g}^T\mathbf{g} + \mathbf{f}^T(\mathbf{K}^T\mathbf{K} + \lambda\mathbf{L}^T\mathbf{L})\mathbf{f} - \mathbf{g}^T\mathbf{K}\mathbf{f} - \mathbf{f}^T\mathbf{K}^T\mathbf{g} \end{aligned} \tag{5.50}$$

If we were to further express the matrix operations in terms of the components of $\mathbf{f}$, $\mathbf{g}$ and $\mathbf{K}$ we would see explicitly that the highest power of $\mathbf{f}_k$ to appear in J is quadratic. Thus following single variable calculus, we expect that setting the gradient of J with respect to the $\mathbf{f}_k$ equal to zero would give a collection of linear equations to be solved for the extremum.[14] This is precisely the case.

To prove this assertion, we introduce some concepts and notation that will prove useful in the coming chapter. Let $\mathbf{y} \in \mathbb{R}^M$ be a function of $\mathbf{x} \in \mathbb{R}^N$. That is $\mathbf{y}_i = \mathbf{y}_i(\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_N)$ for $i = 1, 2, \ldots, M$. The *Jacobian* of $\mathbf{y}$ with respect to $\mathbf{x}$ is the $M \times N$ matrix $\mathbf{J} = \partial\mathbf{y}/\partial\mathbf{x}$ whose (m, n)th element is

$$\mathbf{J}_{m,n} = \frac{\partial \mathbf{y}_m}{\partial \mathbf{x}_n}. \tag{5.51}$$

[14] EXERSICE: Show the extrema is a minimum

With this definition of the Jacobian we show in the exercises at the end of this chapter[15]

$$\frac{\partial}{\partial \mathbf{f}}(\mathbf{K}\mathbf{f}) = \mathbf{A} \quad \text{and} \quad \frac{\partial}{\partial \mathbf{f}}\left(\mathbf{f}^T\mathbf{Q}\mathbf{f}\right) = \mathbf{f}^T(\mathbf{Q}+\mathbf{Q}^T) \underbrace{=}_{\text{if } Q \text{ symmetric}} 2\mathbf{f}^T\mathbf{Q}.$$

Applying these identities to (5.50) and noting that $\mathbf{g}^T\mathbf{K}\mathbf{f} = \mathbf{f}^T\mathbf{K}^T\mathbf{g}$ since both are scalars gives

$$\frac{\partial C}{\partial \mathbf{f}} = 2\mathbf{f}^T(\mathbf{K}^T\mathbf{K} + \lambda\mathbf{L}^T\mathbf{L}) - \mathbf{g}^T\mathbf{K}. \tag{5.52}$$

Setting the transpose of (5.52) to zero then gives (5.49).

The analysis of Tikhonov regularization is both straightforward (using the SVD of $\mathbf{K}$) and interesting in the case where $\mathbf{L} = \mathbf{I}$. The analysis for general $\mathbf{L}$, while still interesting, requires the use of the generalized singular value decomposition which is a bit outside of the scope of the current discussion. We refer the reader to [42]. Moreover for simplicity, let us assume that $\mathbf{K}$ is of full rank with more rows than columns. So, with $\mathbf{L} = \mathbf{I}$ and that the SVD of $\mathbf{K} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ we have

$$\mathbf{K}^T\mathbf{K} = \mathbf{V}\mathbf{\Sigma}_1^2\mathbf{V}^T \quad \text{and} \quad \mathbf{I} = \mathbf{V}\mathbf{V}^T.$$

Substitution into (5.49) gives

$$\begin{aligned}
\hat{\mathbf{f}} &= \left[\mathbf{V}\mathbf{\Sigma}_1^2\mathbf{V}^T + \lambda\mathbf{V}\mathbf{V}^T\right]^{-1}\mathbf{V}\mathbf{\Sigma}\mathbf{U}^T\mathbf{g} \\
&= \left[\mathbf{V}\left(\mathbf{\Sigma}_1^2 + \lambda\mathbf{I}\right)\mathbf{V}^T\right]^{-1}\mathbf{V}\mathbf{\Sigma}\mathbf{U}^T\mathbf{g} \\
&= \mathbf{V}\left(\mathbf{\Sigma}_1^2 + \lambda\mathbf{I}\right)^{-1}\mathbf{\Sigma}_1\mathbf{U}\mathbf{g}.
\end{aligned} \tag{5.53}$$

Expanding (5.53) and taking advantage of the diagonal structure of $\mathbf{\Sigma}_1$ and $\mathbf{I}$ gives

$$\hat{\mathbf{f}} = \sum_k \frac{\sigma_k}{\sigma_k^2 + \lambda}(\mathbf{u}_k^T\mathbf{g})\mathbf{v}_k \tag{5.54}$$

which is precisely the result we obtained using the window function in Example 5.4. Thus we conclude that while Tikhonov regularization with an identity matrix was motivated by a desire to constrain the amplitude of the reconstruction, mathematically, this approach to inversion is identical to a specific instance of spectral windowing.

5.3.5 Parameter Selection

The regularization methods discussed in this chapter all require the specifiction of a parameter in order to generate a reconstruction. Developing useful methods for automatically making this choice has been the topic of considerable work. Here we present a brief overview of the objectives of this work as well as a few of the more commonly used methods for selecting λ. A more detailed discussion including a good deal of analysis can be found in [88, Chpaters 1 and 7].

Methods for selecting a regularization parameter fall into two basic categories. *A priori* approaches assume some knowledge both of the level of noise in the data as well as the nature of the

[15] EXERICES

true object $\mathbf{f}$. For example the so-called *range condition* [88, Section 1.1.2] requires that $\mathbf{f} \in \mathcal{R}(\mathbf{K}^T)$. While such assumptions provide for some very interesting and useful analysis, as our purpose here is a bit more on the practical side, we restrict attention to *a posteriori* parameter selection methods which require only knowledge of the data, including, perhaps, a bound on the intensity of additive noise.

The specification and analysis of parameter selection methods revolves around three ways of quantifying the error in an inversion scheme:

$$\text{Reconstruction error:} \quad \mathbf{e}_\lambda = \quad \hat{\mathbf{f}}_\lambda - \mathbf{f} \tag{5.55}$$

$$\text{Predictive error:} \quad \mathbf{p}_\lambda = \quad \mathbf{K}(\hat{\mathbf{f}}_\lambda - \mathbf{f}) \tag{5.56}$$

$$\text{Residual error:} \quad \mathbf{r}_\lambda = \quad \mathbf{K}\hat{\mathbf{f}}_\lambda - \mathbf{g} \tag{5.57}$$

where $\mathbf{f}$ is the true object and we have made the dependence of the reconstruction on λ explicit. Of the three forms of error, only the residual error can actually be computed. The other two both require knowledge of the true object which, obviously, we do not possess. Nonetheless, it is possible to provide rules for selecting λ which ensure that, even though $\mathbf{e}_\lambda$ and $\mathbf{p}_\lambda$ may not be known as the noise level goes to zero, $\hat{\mathbf{f}}_\lambda$ does in fact go to $\mathbf{f}$.

As a simple example, consider methods such as spectral filtering that take the form of a linear operator, (matrix), $\mathbf{R}_\lambda$, acting on the data. Assuming that the data are described by an additive noise mode, $\mathbf{g} = \mathbf{K}\mathbf{f} + \mathbf{n}$, the reconstruction error is

$$\mathbf{e}_\lambda = (\mathbf{R}_\lambda \mathbf{K} - \mathbf{I})\,\mathbf{f} + \mathbf{R}_\lambda \mathbf{n} \tag{5.58}$$

Thus the error induced by regularization comes from two sources. The first, $\mathbf{e}_\lambda^{trunc} = (\mathbf{R}_\lambda \mathbf{K} - \mathbf{I})\,\mathbf{f}$ is called truncation error and arises because the regularized "inverse" operator is not longer a left inverse of $\mathbf{K}$ as was the case with the pseudo-inverse. The second term $\mathbf{e}_\lambda^{noise} = \mathbf{R}_\lambda \mathbf{n}$ represents the impact of noise amplification by $\mathbf{R}_\lambda$. Thus choosing $\lambda 0$ for linear regularization schemes amounts to trading off these two sources of error.

For spectral filtering methods such as TSVD and Tikhonov regularization with an identity, it is not too difficult to show that $\|\mathbf{e}_\lambda\|_2 \to 0$ as the noise gets small. In this case, we can use the SVD of $\mathbf{K}$ to write the truncation and noise amplification errors as[16]

$$\mathbf{e}_\lambda^{trunc} = \sum_k \left[w_{\lambda,k} - 1\right](\mathbf{v}_k^T \mathbf{f})\mathbf{v}_k \tag{5.59}$$

$$\mathbf{e}_\lambda^{noise} = \sum_k w_{\lambda,k} \frac{1}{\sigma_i}(\mathbf{u}_k^T \mathbf{n})\mathbf{v}_k \tag{5.60}$$

By the requirements of a window function on page 113 , we know that $w_{\lambda,k} \to 1$ as $\lambda \to 0$. Hence in the limit of vanishing regularization parameter, the truncation error will also go to zero. To analyze the impact of vanishing noise, suppose that we know that norm of the error in the data $\|\mathbf{n}\|_2 = \|\mathbf{g} - \mathbf{K}\mathbf{f}\|_2$ is at most δ. For both TSVD and Tikhonov with the identity it is possible to show that[17]

$$\frac{w_{\lambda,k}}{\sigma_i} \leq \frac{1}{\sqrt{\lambda}}$$

[16] EXERCISE: Show these

[17] EXERCISE: Show

so that by judicious use of the triangle inequality as well as the orthonormality of the $\mathbf{u}_k$ and $\mathbf{v}_k$, we can bound (5.60) by

$$\mathbf{e}_\lambda^{noise} \leq \delta/\sqrt{\lambda}. \tag{5.61}$$

So, if we know δ then we can choose $\lambda = \delta^p$ of $0 < p < 2$ and we can guarantee that as the noise goes to zero, so too will $\|\mathbf{e}_\lambda^{noise}\|_2$. A parameter selection method with this property is said to be *convergent*.

As indicated previously, in practice, one cannot determine the reconstruction error. Thus many methods implementable methods for selecting parameters are based on the use of the computable residual error as a provably good proxy for the predictive error. Here we mention four such methods and point the reader to [88, Chapter 7] for many more details.

The Disctrpancy Principle

The *discrepancy principle* [67] states that λ should be chosen so that

$$\frac{1}{M}\|\mathbf{g} - \mathbf{K}\hat{\mathbf{f}}_\lambda\|_2^2 = \nu^2 \tag{5.62}$$

where ν^2 is a bound on the size of the noise. That is, the parameter In the case where the error in the data is additive, white, and Gaussian with variance ν^2, the expected value of the left hand side of (5.62) is just ν^2.

Universal Predictive Risk Estimator

The first approach, Universal Predictive Risk Estimator (UPRE), is based on the idea of choosing λ to minimize

$$\frac{1}{M}\|\mathbf{p}_\lambda\|_2.$$

Again, under the assumption that we are using a linear regularization methods, let us define $\mathbf{A}_\lambda = \mathbf{K}\mathbf{R}_\lambda$ so that we can write

$$\mathbf{p}_\lambda = (\mathbf{A}_\lambda - \mathbf{I})\mathbf{K}\mathbf{f} + \mathbf{A}_\lambda\mathbf{n}.$$

Assuming that the additive noise is a white Gaussian vector with variance ν^2, one can show that

$$E\left[\frac{1}{M}\|\mathbf{p}_\lambda\|_2^2\right] = \frac{1}{M}\|(\mathbf{A}_\lambda - \mathbf{I})\mathbf{K}\mathbf{f}\|_2^2 + \frac{\nu^2}{M}\text{trace}(\mathbf{A}_\lambda^2). \tag{5.63}$$

Now from (5.57) we have

$$\mathbf{r}_\lambda = (\mathbf{A} - \mathbf{I})\,\mathbf{K}\mathbf{f} - (\mathbf{A} - \mathbf{I})\,\mathbf{n}$$

from which one can show

$$E\left[\frac{1}{M}\|\mathbf{p}_\lambda\|_2^2\right] = E\left[\frac{1}{M}\|\mathbf{r}_\lambda\|_2^2\right] + \frac{2\nu^2}{M}\text{trace}(\mathbf{A}_\lambda) - \nu^2 \equiv U(\lambda). \tag{5.64}$$

Eq. (5.64) says that on average, $U(\lambda)$ predicts the value of $\mathbf{p}$. Moreover, given the data, all three terms in $U(\lambda)$ are in fact computable. Thus it is argued that choosing λ to minimize U is a useful

method for minimizing (in an average sense) the predictive risk. Methods for efficiently finding λ in this manner are discussed in [88, Section 7.1.1].[18].

Generalized Cross Validation

The principle of cross validation in the context of regularization parameter selection says that a good paramater is one which is precidtive of unseen data. More concretely, the cross validation functional is

$$CV(\lambda) = \frac{1}{M}\sum_{k=1}^{M}\left(\left[\mathbf{K}\hat{\mathbf{f}}_\lambda^{(k)}\right]_k - \mathbf{g}_k\right) \tag{5.65}$$

where $\hat{\mathbf{f}}_\lambda^{(k)}$ is the estimate of $\mathbf{f}$ based on a data vector whose kth element, $\mathbf{g}_k$, has been removed. Thus each term in (5.65) measures the success in predicting the kth data point from a reconstruction based on the other $M-1$. In [34], the CV functional fails in certain trivial situations such as when $\mathbf{K}$ is diagonal. To address these shortcoming, they recommended the generalized cross validation approach which seeks λ by minimizing

$$GCV(\lambda) = \frac{\frac{1}{M}\|\mathbf{K}\hat{\mathbf{f}}_\lambda - \mathbf{g}\|_2^2}{\left[\frac{1}{M}\text{trace}(\mathbf{A}_\lambda - \mathbf{I})\right]^2}. \tag{5.66}$$

To develop a feeling for the utility of (5.66), let us consider the case where $\mathbf{A}_\lambda$ arises from a TSVD solution to the problem in which case the number of retained singular values, p, is the inverse of the regularization parameter [28, Sextion 4.5]. Using the singular value decomposition of $\mathbf{K}$, we can show that the matrix in the denominator of $GCV(p)$ is[19]

$$\mathbf{A}_p = \mathbf{U}\begin{bmatrix}\mathbf{I}_p & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}\mathbf{U}^T$$

so that $\frac{1}{M}\text{trace}(\mathbf{I}-\mathbf{A}_p) = 1 - p/M$. Now, let us examine the numerator of $GCV(p)$. Here we have

$$\mathbf{K}\hat{\mathbf{f}}_p - \mathbf{g} = (\mathbf{A}_p - \mathbf{I})\mathbf{K}\mathbf{f} + (\mathbf{A}_p - \mathbf{I})\mathbf{n}.$$

Analysis of this proceeds along the same lines as that of the GCV numerator (to simplify $\mathbf{I}-\mathbf{A}_p$) as well as UPRE (to analyze the expected value of $\|\mathbf{g} - \mathbf{K}\hat{\mathbf{f}}_p\|_2^2$) to arrive at

$$E\left[\|\mathbf{K}\hat{\mathbf{f}}_p - \mathbf{g}\|_2^2\right] = \|(\mathbf{A}_p - \mathbf{I})\mathbf{K}\mathbf{f}\|_2^2 + M\nu^2\left(1 - \frac{p}{M}\right). \tag{5.67}$$

Thus, on average the expected value of $GCV(p)$ is

$$E\left[GCV(p)\right] = \frac{1}{\left(1-\frac{p}{M}\right)^2}\left[\|(\mathbf{A}_p - \mathbf{I})\mathbf{K}\mathbf{f}\|_2^2 + M\nu^2\left(1 - \frac{p}{M}\right)\right] \tag{5.68}$$

Because we are interested in minimizing $GCV(p)$, the monotonically increasing factor $(1-p/M)^{-2}$ can be neglected. As a function of p, the first term in the brackets in (5.68) is decreasing while the

[18] EXERCISE: Implementation of UPRE

[19] EXERCISE: Fill in the gaps

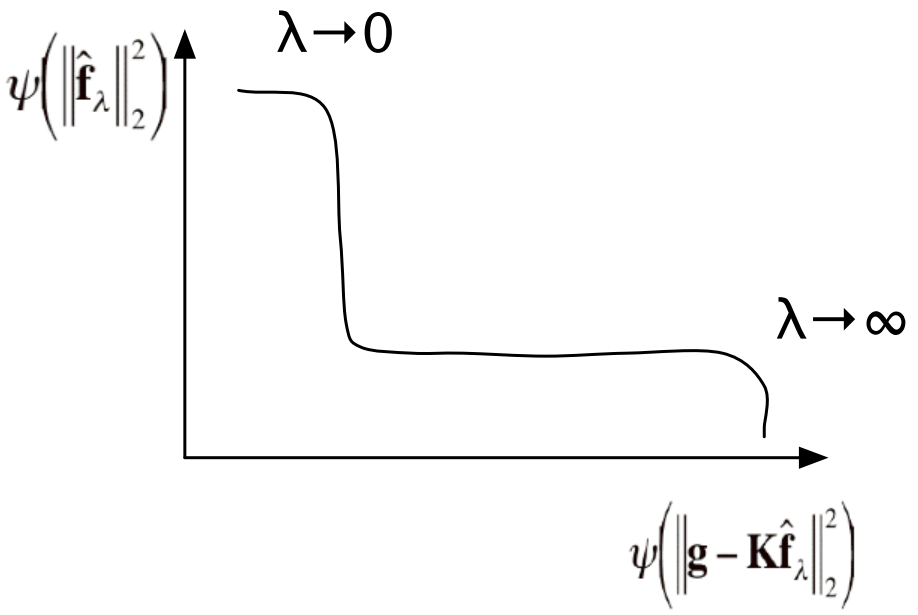


Figure 5.25: Model L Curve

second is increasing. So, one would expect that at least on average, there should be a value for p where the sum of these competing terms is minimum. This can be taken a bit further assuming that the minimizing value of GCV occurs for $p \ll M$. In this case bringing $(1 - p/M)^{-2}$ into the brackets and making the approximations

$$\frac{1}{\left(1-\frac{p}{M}\right)^2}\|(\mathbf{A}_p - \mathbf{I})\mathbf{K}\mathbf{f}\|_2^2 \approx \|(\mathbf{A}_p - \mathbf{I})\mathbf{K}\mathbf{f}\|_2^2$$

$$\frac{M\nu^2}{1-\frac{p}{M}} \approx M\nu^2\left(1+\frac{p}{M}\right)$$

we see that $GCV(p)$ is quite similar to (5.63), the (unknown) predictive risk. Thus, we conclude that minimizing GCV may be a reasonable method for selecting a regularization parameter.

The L-Curve

The last approach to parameter selection we wish to cover here is known as the L-Curve and consists of a plot of $\psi(\|\hat{\mathbf{f}}_\lambda\|_2^2)$ versus $\psi(\|\mathbf{g} - \mathbf{K}\hat{\mathbf{f}}_\lambda\|)$ as λ is varied. The function $\psi(x)$ sets the scale of the graph and is typically $\log(x)$ although $\sqrt{x}$ has also been suggested. The name of this method is motivated by the best-case shape of the curve. Consider the case of Tikhonov regularization. As is shown in Fig. 5.25, when the problem is under-regularized, $\lambda \to 0$, one expects that the norm of the reconstruction will be large due to the artifacts, but the fit to the data will be quite good. As $\lambda \to \infty$, (5.46) says that the fit to the data will be poor since the error term gets little weight,

while the norm of the object will be small as this is dominant term in the minimization. Moreover in [41, 42], it is argued that in these extreme cases, as λ is varied, the L-curve will be vertical and horizontal respectively. Thus, heuristically at least, there must exist a corner where the two forces are in some balance. The value of the parameter at this corner then is the one chosen by the L-curve method. The L-curve has also been used in the context of TSVD where the corner is found by interpolating the individual points along the curve generated as the number of retained singular values is changed. Mathematically,the corner is typically defined as the point of maximum curvature along the L-curve. Algorithmic methods for locating this point are discussed in [4,40,42].

One interesting feature of the L-curve is the discrepancy between its theoretical properties and its performance in practice. As is discussed in some detail in [28, Section 4.5], there exists a body of research indicating that the L-curve will work poorly in certain limiting cases. While the work in [29] points to the necessity of using $\psi = \log$, the failure of the L curve as documented in [39, 86] are a bit more theoretically problematic. The analysis in [86] indicated that the value of the corner parameter will not go to zero as the noise vanishes (hence the method is termed *non-convergent*) while the results in [39] indicate that λ will in fact go to zero too rapidly. While the disparity arises from differences in the underlying model problems being studied, the results clearly indicate that in the limit of low noise, the L-curve may not provide the best approach to choosing λ. In spite of these theoretical results, the L-curve approach has received considerable attention even since the mid-1990's when [39, 86] appeared. Work is ongoing in terms of efficiently locating the corner of the L-curve, generalizing the notion of the L-curve to multiple parameters, and in applying the methods to a range of applications.

The primary conclusion to be drawn from this is that the utility of the L-curve (really all of the methods discussed here) as an automatic tool for selecting the "best" value of a regularization parameter is far from a settled question. While methods of regularization such as spectral windowing, the TSVD, and various variational approaches are well defined, the issue of parameter selection is one dominated by heuristic approaches whose behaviors, while analyzable in the certain limits, are far more difficult to predict in practice. Hence for the time being at least, proper parameter choice still requires a level of human interaction with the inverse routine.

5.3.6 Examples

5.4 Semi-discrete Linear Inverse Problems

- Probably save for next time class is taught.
- Add dual basis ideas to math background chapter as well as numerical differentiation problem
- Problem formulation: finite data but want to recover continuous function.
- Topics to cover: Form of pseudo-inverse, spectral shaping, Tikhonov, change of function space for these problems. Numerical differentiation example continued. Link to fully discrete problem.

5.5 Exercises

5.1 In this problem we investigate inverses in the full rank case.

1. Consider a discrete convolutional problem corresponding to a first kind Fredholm equation:
$$y(i) = \sum_j h(i-j)x(j) \tag{5.69}$$
As you showed in Problem Set 1, we can represent such a problem in the form $y = Cx$ where C is an $N \times N$ circulant matrix whose nonzero entries are given by $h(i)$. Often times a change of basis can make a problem easier. Suppose we make the change of basis $X = Fx$ and $Y = Fy$, where F is the DFT matrix. What is the corresponding problem in the transformed space? Assuming that $\text{rank}(C) = N$, show that we may write the solution as:
$$x = \frac{1}{N} F^H \text{diag}[1/H(k)] F y \tag{5.70}$$
where $H(k)$ are the DFT elements of $h(i)$ and A^H is the conjugate transpose of the matrix A. Interpret this procedure in the frequency domain.
2. The above change of basis will not work for general linear operators since they do not obey a "Fourier-convolution theorem". All is not lost, however. Consider the general linear problem $y = Ax$ where A is an $N \times N$ matrix of $\text{rank}(A) = N$ with no particular structure. Let the SVD of A be given by $A = USV^T$. What changes of bases would lead to a corresponding diagonal system? Give an expression for the solution x in terms of the SVD that is analogous to (3.38). Provide an equivalent "Frequency domain" interpretation of this result. What are the corresponding generalized Fourier coefficients corresponding to A. What is the corresponding generalized Fourier transform operator in this case? In what major way does it differ from the standard Fourier operator or, equivalently, in what major way does this change of basis differ from that used in part (a)? What computational challenges does this present?

5.2 For this problem, you will need to retrieve the file **illposed.mat.gz** from the course web site. In this MATLAB file there is a matrix K for the forward problem of interest here, as well three input output vector pairs, f_i and g_i $i = 1, 2, 3$ respectively. Thus, the problem of interest here is the recovery of f from g where the two are related by the matrix-vector equation $g = Kf$.

1. As discussed in class, a problem is considered ill-posed due to difficulties which include the issues of the existence and uniqueness of the solution. Based on your analysis of K, which of these issues are relevant here? Explain.
2. The third characteristic of an ill-posed problem is an undue sensitivity of the solution to small changes in the data. One way of quantifying this notion wis by looking at the quantity
$$\frac{\|f_1 - f_2\|/\|f_1\|}{\|g_1 - g_2\|/\|g_1\|} \tag{5.71}$$

defined for two input/output pairs. If (5.71) is large ($\gg 1$) then the problem is rather sensitive to small changes in the data. Based on the data supplied in **illposed.mat.gz**, can this problem be ill-posed? That is, do there exist "close" outputs which correspond to significantly different inputs? If so, provide a quantitative discussion describing those characteristics of these pairs which cause (5.71) to be large?

3. Is it the case that all input/output pairs lead to (5.71) being $\gg 1$? If so, explain why. Otherwise, find two input/output pairs for which (5.71) is ≈ 1 and explain how you constructed these signals.
4. Consider the case where the matrix K is invertible. It is claimed that under this condition one can always exactly recover an object by merely applying K^{-1} to the data. Why is this true only in theory? Using the K from **illposed.mat.gz**, construct a problem for which the application of K^{-1} produces an object which is much, much different from the one intended. Explain your construction.

5.3 Here we consider the inverse problem of numerical differentiation, where one is given observations of the running integral of a function and attempts to recover the original function. For this problem the data $g(x)$ are related to the function of interest, $f(x)$, through:

$$g(x) = \int_0^x f(y)dy \quad 0 \leq x \leq 1 \tag{5.72}$$

where $f(x) = 0$ for $x < 0$. In this problem we investigate the effect of perturbations on the solution.

1. It is clear that the exact solution to the continuous problem is given by:

$$f(x) = \frac{d}{dx}g(x) \tag{5.73}$$

Now suppose instead of the exact $g(x)$ we observe a slightly perturbed version $g_p(x)$, given by:

$$g_p(x) = g(x) + \epsilon \sin(\omega x) \tag{5.74}$$

i.e. there is a small high-frequency "noise" superimposed on the signal. What is the corresponding solution $f_p(x)$ based on (5.73)?

2. What happens to $||g(x) - g_p(x)||_2$ (i.e. the energy in the observation error) as $\epsilon \to 0$ and $\omega \to \infty$? What happens to $||f(x) - f_p(x)||_2$ (i.e. the corresponding energy in the reconstruction error) as $\epsilon \to 0$ and $\omega \to \infty$? What does this suggest about the stability of the noise-free, infinite-dimensional case?
3. Now we want to examine a fully discrete form of this problem. All discretization are to obtained using a Galerkin approach with flat-top basis functions and impulse testing functions. That is, $f(x)$ is expanded as

$$f(x) = \sum_{n=1}^{N} f_n \phi_n(x)$$

with

$$\phi_x(x) = \begin{cases} 1 & \frac{(n-1)}{N} \leq t \leq j\frac{1}{N} \\ 0 & \text{else} \end{cases}$$

and the data vector is composed of N equally spaced samples of $g(x)$ for $x \in [0, 1]$. The final result of this discretization is a matrix-vector relationship of the form $g = Af$ where g is the vector of sampled values of $g(x)$, f is the vector of expansion coefficients, and A is the matrix obtained by discretizing the kernel in (5.72). All parts to this problem are to be carried out for $N \in \{10,\ 100,\ 300,\ 500\}$.

For the case where $f(x) = u(x)$ (the unit step), create a perturbed version g_p of the discrete observation vector g, according to the following formula for the k-th element:

$$(g_p)_k = g_k + \frac{(-1)^k}{\sqrt{N}} \tag{5.75}$$

Note that this is a discrete version of the perturbation considered in part (a). Generate plots of the perturbed observation $(g_p)_k$ vs index k for each N. What happens to the perturbation as N increases; i.e. what, if anything, does g_p tend to.

4. Generate plots of the corresponding discrete estimates $\widehat{x}_p = A^{-1}y_p$ versus index for each N. Does $\widehat{x}_p$ approach $\widehat{x}$?
5. Finally, summarize these results by making plots of the percentage errors in y_p and $\widehat{x}_p$ as a function of N. These are defined as:

$$\%\text{ error in } y = 100\frac{||y - y_p||_2}{||y||_2}, \quad \%\text{ error in } \widehat{x}_p = 100\frac{||\widehat{x} - \widehat{x}_p||_2}{||\widehat{x}||_2} \tag{5.76}$$

5.4 This problem investigates a discrete Picard condition. For this problem you will need access to routines from the Regularization Toolbox. If this is not already installed on your system there are links to it on the class web site.

1. As discussed in class, the generalized solution to an inverse problem is given by:

$$x^+ = \sum_{i=1}^{N} u_i \frac{\langle v_i, y\rangle}{\alpha_i}$$

Since the singular values decrease as i increases, this solution will be "stable" only if the terms $|\langle v_i, y\rangle|$ decay faster than the singular values α_i as $i \to N$. This condition has been proposed as a discrete Picard condition. Why is the term "Picard condition" an appropriate one?

2. For the discrete differentiation problem use the routine picard.m to plot the unperturbed quantities $|\langle v_i, y\rangle|$, α_i, and $|\langle v_i, y\rangle|/\alpha_i$ versus i for $N = [10,\ 100,\ 300,\ 500]$. Repeat the plots using the perturbed data sequences. Using the plots explain the difference in behavior between reconstructions based on the unperturbed and perturbed data. Notes: The routine picard.m takes as input U, s, and y, where s is a *vector* of the *non-zero* singular values and U are the corresponding singular functions. These can be conveniently generated in MATLAB using e.g. SVD(C,0).

3. In the discrete differentiation problem the perturbation introduced was deterministic, yet the same difficulties occur in the stochastic case. Suppose we are given a discrete inverse problem of the form $y = Cx = \sum_i v_i \alpha_i u_i^T x$, y is the exact observation corresponding to x. In real applications a more realistic model is that we instead observe a noisy version of y given by:
$$y_n = y + w \quad w \sim N(0, \sigma^2 I) \tag{5.77}$$
i.e., w is a vector of independent identically distributed noise samples. What is $E(|v_i^T w|)$, the expected value of the magnitude of the generalized Fourier coefficients corresponding to the noise? Even if the unperturbed problem behaves well, what problems does your answer suggest, given that the α_i always decay towards zero?
4. Confirm the insight developed in part (c) by generating a noisy version of the discrete observation for the differentiation problem for $N = 300$ according to the formula (5.77) with $\sigma = 0.1$. Another difficulty is that, for most problems it turns out that the corresponding singular functions v_i become more oscillatory as i increases (just as in standard Fourier analysis). Thus the functions getting the largest weighting in the generalized reconstruction are those that are more "noise like". Lets see if this phenomenon is true for the differentiation problem. Plot v_i for $i = [1, 10, 100, 300]$.

5.5 In this problem we examine the fully discrete tomography problem. We will use the exact phantom in small_phantom.mat as our starting point – call it x. In this problem, we want to consider two tomographic scenarios

LA The "limited-angle" problem where there angular coverage is far less than the full 2π radians. For the problem below, you should generate C_{la} matrices corresponding to 16 equally spaced angles in [0, 90).

SD For the sparse data problem, one can measure only a few projections over the entire 2π radian field of view. The matrices C_{sd} for this problem should correspond to 8 equally spaced angles in [0, 180).

In all cases use the routines you have developed on previous problem sets to generate the matrices of interest.

1. For both problems, use your C's to generate clean projection data y. In addition, generate a noisy version y_n of the observation y according to the formula (5.77) with $\sigma^2 = 9$. What is the SNR for these cases?
2. For each problem, find the SVD of C and use the routine picard.m to plot the noise free quantities $|\langle v_i, y\rangle|$, α_i, and $|\langle v_i, y\rangle|/\alpha_i$ versus i. Repeat the plots using the perturbed data y_n. In addition, plot the corresponding singular functions u_i for $i = [1, 90, 180, 250]$ again *as images*. Given these results, what type of reconstructions do you expect? What are the condition numbers of C?
3. Now examine the generalized solutions together with our friend the FBP solution for both **LA** and **SD**. Find the generalized solutions corresponding to both the noiseless data (i.e. $C^+ y$) and the noisy data (i.e. $C^+ y_n$) using the SVD and the also find the FBP solution for both cases. Plot all these solutions, compare, and comment.

5.6 In class one day, I mentioned that the least squares solution to an overdetermined, full rank problem was just one way one could consider finding a "solution" to a problem which technically had no solution. Here we want to look at another. The setting is as follows. Suppose that we had a linear inverse problem of the form $g = Uf$ where g is a length n vector and U is an unknown $n \times n$ orthonormal matrix. In addition to these data, we also know that f lies in the linear span of a collection of linearly independent vectors $\{a_1, a_2, \cdots a_m\}$ with $m < n$. I am curious about developing methods for estimating f which make use of these two pieces of information.

1. For the case where $a_1 = [1, 0]^T$ and $g = [5, -3]^T$ use the projection-onto-the-range-of-A interpretation of the pseudo-inverse as the basis for finding an f that solves the problem. Explain in detail why your solution in not unique.
2. Generalize the above discussion to the case where $a_1 = [1, 0, 0]^T$ and $a_2 = [0, 1, 0]^T$. and g is any length three vector. Specifically, sketch the nonuniqueness region in the range of A.
3. Here is a suggestion for making the solution more unique: Select that point in the region of non-uniqueness which is closest to the pseudo-inverse solution.
   (a) Why might this be a good idea? Why might it be a bad idea?
   (b) For the same A as in the previous part of this problem, explain in detail why this strategy will not *always* yield a unique solution?
   (c) For those cases where the solution will be unique, find the analytical solution to the inverse problem again for the 3×2 A specified previously.

5.7 Now let us return to the Born inverse scattering problem which was the subject of Problem 2.2. Here we want to explore the performance of different inverse methods for a variety of sensing configurations:

Configuration 1: 20 sources equally spaced on left side and 20 receivers equally spaced on the right

Configuration 2: 5 sources/receiver pairs equally spaced on each of the 4 sides of the region.

In each case we have a total of 800 data points (400 real and 400 imaginary). The remaining parameters for the problem should be set as follows: $N_y = N_z = 30$, $Z = Y = x_0 = 1$, $d = 0.1$, $\delta = 0.01$, $k_0 = 25 + \sqrt{-1}$.

For each of these configurations, you are to examine the performance of the pseudo-inverse and TSVD (you pick the truncation level, but explain how) on the recovery of the two object function `f1` and `f2` generated from the following Matlab code:

```
f1 = zeros(30);
f1(8:10,12:20) = 1;
f1(20:25,20:25) = 2;

f2 = zeros(30);
```

```
for i = 1:30
  for j = 1:30
    tmp1 = exp(-( ((i-9)^2/10) + ((j-15)^2/50)));
    tmp2 = exp(-((i-23)^2+(j-23)^2)/50);
    f2(i,j) = tmp1+2*tmp2;
  end
end
```

Analyze the results using both noise-free data as well as data corrupted by a zero mean, additive white Gaussian noise vector whose standard deviation, σ is set such that the SNR=30dB where SNR is defined as

$$SNR = 10 \log_{10} \frac{\|Af\|_2^2}{\sigma^2 N}$$

where N is the length of the vector Af.

5.8 Least squares problem where there is a prior mean on $\mathbf{f}$.

Chapter 6

Numerical Methods for Nonlinear Inverse Problems

A non-linear inverse problem is one where $\hat{\mathbf{f}}$ cannot be written as a linear (or even affine) function of the data, $\mathbf{g}$. A careful examination of the inverse methods discussed in the last two chapters will show that three conditions are required to obtain a linear inverse method. First, the physics of the problems must either be linear (convolution or X-ray tomography) or well approximated as linear (the Born approximation). Given a linear forward model, the unknowns must be obtained as the result of a basis expansion type of method as in (3.44). That is, we must be inverting for pixel-type of quantities. Finally, should a variational type of regularization be employed as in § 5.3.3, $\rho(\mathbf{f})$ must be quadratic in the unknown pixel values. If any of these conditions fail to be met then, at least in the variational context, the resulting optimization problem will not possess a linear solution as in (5.48). To be clear, even if the physics are linear, should we find it useful to invert for quantities other than pixel-type expansion coefficients or should be choose to use pixels but regularize in some way other than with a two norm penalty, we will be faced with the need to solve a nonlinear inverse problem.

In fact, these possibilities are quite likely both in research as well as more applied settings. In terms of nonlinearity of the physics, one may not possess sufficient information regarding the background to build a valid Born model. Thus either the full scattering physics [17,36,38,61,63,65] or at least a higher order approximation may need to be used in an inverse scheme. Similarly, it is well known that quadratic regularization tends to produce imagery in which sharp discontinuities such as edges are not accurately recovered. Thus, there has been quite a bit of work over the past decade exploring the use of non-quadratic regularization methods designed specifically to enhance these important image features [3, 13, 31, 87]. Finally, the inversion for quantities other than pixels, specifically for parameters related to the shape of an unknown scattering object, has a long and highly interesting history in the mathematics and mathematical physics literature where it is known as the inverse obstacle problem [44, 51, 81] as well as the signal and image processing communities [25, 32, 54, 59, 60, 62, 78].

As was the case with linear inverse problems, there are two basic approaches for solving nonlinear variants: analytical and numerical. Unlike the linear case here we start with the numerical and present the analytical methods subsequently. This unfortunate pedagogical asymmetry results

from the underlying differences in which these two classes of problems are solved in practice. While at the current time, the vast majority of inverse methods employed in any regular practice are linear, within the context of nonlinear inversion, numerical methods based on variational principles are far closer to being used in practice than the analytical alternatives. A number of factors are at work here. First, the numerical methods for nonlinear problems are in many ways easy extensions of the ideas presented in Chapter 5 for linear problems. This is not the case for the analytical techniques which generally require far different and far more sophisticated mathematical methods than are encountered in the study of filtered backprojection or filtered backpropagation. Second, for many problems such as X-ray tomography and deconvolution where the physics are exactly linear, the introduction of nonlinearity via regularization comes directly in the context of a variational formulation of the problem so that numerical approaches are in a sense the most natural option. Finally, for many inverse obstacle type of problems, there exist no known analytical solution methods. Hence a numerical formulation is the only viable option. Thus, as we are motivated in this manuscript primarily by the desire to provide exposition on practically useful inversion schemes, we start our discussion of nonlinear inverse problems with numerical approaches arising from variational formulations.

Fundamentally then, in this chapter a nonlinear inverse problems is equivalent to the solution of a non-quadratic numerical optimization problem. Thus, we begin with a review of some basic methods in numerical optimization concentrating specifically on techniques for solving a class of problems, non-linear least squares, which shall be encountered repeatedly in subsequent discussions. Armed with these tools we shall then discuss in some detail the most straightforward and basic forms of the three classes of problems described at the start of this section: non-quadratic regularization for problems with linear physical models, inverse obstacle problems again with linear physical models, and finally, inverse problems where the physics are nonlinear, the unknowns are pixels, and the regularization is arbitrary.

6.1 A Review of Optimization Theory and Algorithms

The issue of optimization methods in the context of inversion is not as straightforward as one might suspect. It is possible (and quite frequently done) to use the ideas presented in this section and elaborated upon in a number of more complete sources such as [45,70,88] to build one's own suite of computational tools for solving the optimization problem encountered in a given inversion problem. While such an approach may seem attractive, unless one has a background in the well developed field of numerical methods for optimization, it is unlikely that the resulting code will be as efficient, robust, or capable of solving large scale problems as professionally developed tools by companies such as the Mathworks, IMSL, and the Numerical Algorithms Group.

The precise mix of commercial and self-written code is highly problem dependent. One downside to using commercial tools, in addition to their cost, is their "black box" nature. That is, one does not have complete control over the optimization approach and hence cannot easily change the internal structure to exploit any particular structure which may exists for the given inverse problem. On the other hand, naive implementation of methods presented here may well fail to converge (or converge at an unacceptably slow rate) to a local minimum of the cost function. Such is the case for edge preserving regularization problems which we discuss in §XXX.. As is explained in great

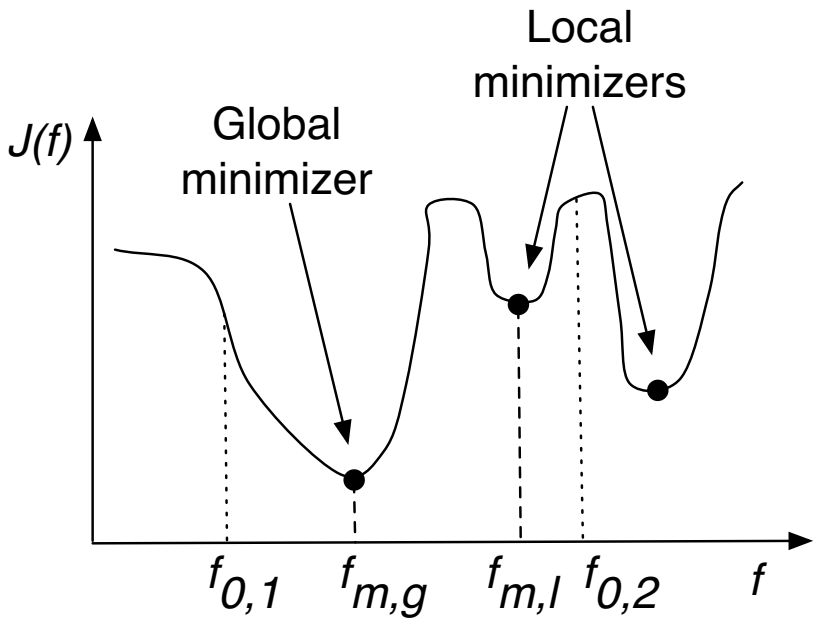


Figure 6.1: Global and local minimizers of a cost function

detail in [88, Chapter 8], the full benefit of this powerful regularization method may well require the use of optimization methods far more sophisticated than those covered in this manuscript. Ultimately, there is a certain amount of iteration among computational tools that must be done for any problems to understand the numerical intricacies and hence choose the most appropriate tools.

Finally, there exist a number of well known specialized optimization methods for solving particular problems which differ from the off-the-shelf types of tools considered in this section. As we encounter the relevant problems later in this chapter we shall discuss these specific methods individually.

6.1.1 General Unconstrained Problems

In this manuscript we are concerned with the solution to unconstrained, non-quadratic optimization problems of the form

$$\hat{\mathbf{f}} = \arg\min_{\mathbf{f}} J(\mathbf{f}) \tag{6.1}$$

where $\mathbf{f} \in \mathbb{R}^N$ and the cost function, J, is a mapping from $\mathbb{R}^N$ to $\mathbb{R}$ and is assumed to be continuously differentiable in the elements of $\mathbf{f}$. There are two classes of minimizers that are generally considered for J. A *global* minimizer is any $\hat{\mathbf{f}}$ such that for all other $\mathbf{f}$, $J(\hat{\mathbf{f}}) \leq J(\mathbf{f})$. A *local* minimizer is any $\hat{\mathbf{f}}$ such that there is a δ for which $\|\hat{\mathbf{f}} - \mathbf{f}\| \leq \delta$ implies $J(\hat{\mathbf{f}}) \leq J(\mathbf{f})$. In the 1D case the difference between these two concepts is illustrated in Fig. 6.1.

The methods used to find a minimum of (6.1) that we consider here all take the form shown in Algorithm 1. The user provides an initial guess as to $\hat{\mathbf{f}}$ which is iteratively refined to produce increasingly more accurate approximations to a local minimum of J. Specifically, the local minimum to which the algorithm converges is one within the "basin of attraction" of the initial guess. Again referring to Fig. 6.1, if we start at $\mathbf{f}_{0,1}$, then the global minimum, $\mathbf{f}_{m,g}$ will be found; however if the initial estimate is $\mathbf{f}_{0,2}$, the result will be a local minimizer of the cost function. The differences among the methods we consider arise from the manner in which update directions and step sizes are calculated as well as the convergence criteria used.

Algorithm 1 Generic Nonlinear Optimization Algorithm

$n = 0$
$\mathbf{f}^{(n)} = \mathbf{f}_0$ {$\mathbf{f}_0$ = user supplied initial guess}
repeat
  Compute $\mathbf{d}^{(n)}$, an update direction for $\mathbf{f}$
  Compute $\tau^{(n)}$, a stepsize indicating how far we move in the direction $\mathbf{d}^{(n)}$
  $\mathbf{f}^{(n+1)} = \mathbf{f}^{(n)} + \tau^{(n)}\mathbf{d}^{(n)}$
  $n = n + 1$
until Convergence
$\hat{\mathbf{f}} = \mathbf{f}^{(n)}$

The assumption that J is continuously differentiable in the components of $\mathbf{f}$ significantly simplifies the mathematical criterion that must be met for some $\mathbf{f}$ to be a local minimizer of J. Specifically, [70, Theorem 2.2] says that necessary conditions for $\hat{\mathbf{f}}$ to be a local minimizer of J are $\nabla J(\hat{\mathbf{f}}) = \mathbf{0}$ where $\nabla J(\mathbf{f})$ is the length N column vector whose ith element is $\partial J/\partial \mathbf{f}_i$. The numerical methods for finding $\hat{\mathbf{f}}$ are all geared to ensuring these conditions are met though their choice of $\mathbf{d}^{(n)}$ and $\tau^{(n)}$ in Alg. 1.

Two choices are established the analysis of the Taylor series of J taken about $\hat{\mathbf{f}}^{(n)}$. Dropping the explicit dependence on the iteration number n yields to second order

$$J(\mathbf{f} + \tau\mathbf{d}) \approx J(\mathbf{f}) + \tau\mathbf{d}^T\nabla J(\mathbf{f}) + \frac{1}{2}\tau^2\mathbf{d}^T\nabla^2 J(\mathbf{f})\mathbf{d} \tag{6.2}$$

where $\nabla^2 J(\mathbf{f})$ is the $N \times N$ Hessian matrix whose (i,j)-th element is $\frac{\partial^2 J}{\partial \mathbf{f}_i \partial \mathbf{f}_j}$. The *steepest decent* method for choosing $\mathbf{d}$ is obtained by assuming the second order term can be ignored and searching for that unit norm $\mathbf{d}$ which yields the largest decrease in J. It is readily verified that the solution to this problem is to take

$$\mathbf{d} = -\frac{\nabla J(\mathbf{f})}{\|\nabla J(\mathbf{f})\|}. \tag{6.3}$$

Alternatively, *Newton's* method searches for a decent direction which minimizes the full right hand side of (6.2). The result is

$$\mathbf{d} = -\left[\nabla^2 J(\mathbf{f})\right]^{-1} \nabla J(\mathbf{f}). \tag{6.4}$$

It turns out that in this case, the step length can be taken as $\tau = 1$ [70, Section 2.2].

A key difference between steepest decent and Newton's method is the need to compute second derivative information for the later. As we shall see in coming sections, it is typically the case that

obtaining the gradient is a rather cumbersome procedure. Exact evaluation of the Hessian is thus not frequently an option much less determining its inverse. The primary benefit of the Newton method (and hence the extra work of calculating the Hessian) is the improvement in convergence relative to steepest decent. Subject to certain technical conditions and proper choice of the step length (see below), it can be shown that for steepest decent

$$\|\mathbf{f}^{(n+1)} - \hat{\mathbf{f}}\| \leq \|\mathbf{f}^{(n)} - \hat{\mathbf{f}}\|$$

while for Newton's method

$$\|\mathbf{f}^{(n+1)} - \hat{\mathbf{f}}\| \leq \|\mathbf{f}^{(n)} - \hat{\mathbf{f}}\|^2.$$

Thus, steepest decent converges at a linear rate while Newton does so quadratically. The difference can be substantial especially when one is close to the minimum.

Quasi-Newton methods represent one method for improving convergence without requiring additional derivatives. The basic idea underlying this class of techniques is to replace Hessian, $\nabla^2 J(\mathbf{f}^{(n)})$ in (6.4), with a more easily computed matrix $\mathbf{B}^{(n)}$. The most widely used technique in this class is the Broyden-Fletcher-Goldfarb-Shanno (BFGS). This approach directly calculates $\mathbf{H}^{(n)} \equiv \left[\mathbf{B}^{(n)}\right]^{-1}$ thereby avoiding the need for a matrix inversion or equivalently a linear system solve at each iteration. Moreover, it yields superlinear (although not quadratic) convergence [70, Section 8.4]. Defining:

$$\mathbf{s}^{(n)} = \mathbf{f}^{(n+1)} - \mathbf{f}^{(n)} \quad \text{and} \quad \mathbf{y}^{(n)} = \nabla J\left(\mathbf{f}^{(n+1)}\right) - \nabla J(\left(\mathbf{f}^{(n)}\right)$$

the BFGS method is

$$\mathbf{H}^{(n+1)} = \mathbf{H}^{(n)} - \left(\mathbf{I} - \frac{\mathbf{s}^{(n)}\mathbf{y}^{(n)^T}}{\mathbf{y}^{(n)^T}\mathbf{s}^{(n)}}\right)\mathbf{H}^{(n)}\left(\mathbf{I} - \frac{\mathbf{y}^{(n)}\mathbf{s}^{(n)^T}}{\mathbf{y}^{(n)^T}\mathbf{s}^{(n)}}\right) + \frac{\mathbf{s}^{(n)}\mathbf{s}^{(n)^T}}{\mathbf{y}^{(n)^T}\mathbf{s}^{(n)}} \tag{6.5}$$

Finally, the *non-linear conjugate gradient* approach, is based on a substantially different theoretical foundation than steepest decent, Newton, or quasi-Newton. While that theory is a bit beyond the scope of this manuscript (see [70, Chapter 5]), the algorithm takes the form, slightly different from Alg. 1 and is shown in Alg. 2. Two common methods are used for computing $\beta(n)$:

Algorithm 2 Nonlinear Conjugate Gradient

$n = 0$
$\mathbf{f}^{(n)} = \mathbf{f}_0$ {$\mathbf{f}_0$ = user supplied initial guess}
$\mathbf{d}^{(n)} = \nabla J(\mathbf{f}^{(n)})$
repeat
  Compute $\tau^{(n)}$
  $\mathbf{f}^{(n+1)} = \mathbf{f}^{(n)} + \tau^{(n)}\mathbf{d}^{(n)}$
  Compute $\beta^{(n)}$
  $\mathbf{d}^{(n+1)} = -\nabla J(\mathbf{f}^{(n+1)}) + \beta^{(n)}\mathbf{d}^{(n)}$
  $n = n + 1$
until Convergence
$\hat{\mathbf{f}} = \mathbf{f}^{(n)}$

$$\text{Fletcher-Reeves:} \quad \beta^{(n)} = \frac{-\|\nabla J\left(\mathbf{f}^{(n+1)}\right)\|_2^2}{\|\nabla J\left(\mathbf{f}^{(n)}\right)\|_2^2} \tag{6.6}$$

$$\text{Polak-Ribière:} \quad \beta^{(n)} = \frac{-\left[\nabla J\left(\mathbf{f}^{(n+1)}\right)\right]^T\left[\nabla J\left(\mathbf{f}^{(n+1)}\right) - \nabla J\left(\mathbf{f}^{(n)}\right)\right]}{\|\nabla J(\mathbf{f}^{(n)})\|_2^2} \tag{6.7}$$

with the Polak-Ribière preferred in practice.

Both the steepest decent and the nonlinear conjugate gradient methods require the determination of a step-length parameter, $\tau^{(n)}$. In theory, this quantity should be chosen to minimize the cost function in the direction $\mathbf{d}^{(n)}$:

$$\tau^{(n)} = \arg\min_{\tau} J\left(\mathbf{f}^{(n)} + \tau\mathbf{d}^{(n)}\right). \tag{6.8}$$

In practice, performing this added optimization step to a high degree of accuracy is both computationally costly and not really necessary to guarantee the convergence of the overall algorithm. Thus a number of *inexact linear search* methods have been developed with proven convergence properties. We refer the reader to [70, Chapter 3] and [88, Section 3.4] for detailed discussions of these methods.

6.1.2 Nonlinear Least Squares Problems

In many inverse problems, the cost function takes a particular form leading to *non-linear least squares* optimization methods. Such problems arise if we can write

$$J(\mathbf{f}) = \frac{1}{2}\mathbf{e}(\mathbf{f})^T\mathbf{e}(\mathbf{f}). \tag{6.9}$$

For example, $\mathbf{e}_k(\mathbf{f})$ could represent the difference between a measured datum and the prediction of a model given $\mathbf{f}$, *i.e.* $\mathbf{e}_k(\mathbf{f}) = \mathbf{g}_k - \mathbf{h}_k(\mathbf{f})$ although as we shall see the use of regularizers leader to other structure for $\mathbf{e}$.

The particular structure in (6.9) is reflected in the required gradient calculations for methods such as steepest decent and leads to a pair of alternate quasi-Newton methods. Defining the Jacobian, $\mathbf{J}$, for this problem as the $M \times N$ whose (i,j)-0th element is

$$\mathbf{J}(\mathbf{f})_{i,j} = \frac{\partial \mathbf{e}_i(\mathbf{f})}{\partial \mathbf{f}_k}$$

direct calculation shows

$$\nabla J(\mathbf{f}) = \mathbf{J}(\mathbf{f})\mathbf{e}(\mathbf{f}) \tag{6.10}$$

$$\nabla^2 J(\mathbf{f}) = \mathbf{J}(\mathbf{f})^T\mathbf{J}(\mathbf{f}) + \sum_{k=1}^{M} \mathbf{e}_k(\mathbf{f})\nabla^2\mathbf{e}_k(\mathbf{f}). \tag{6.11}$$

Eq. (6.11) shows that a portion of the Hessian can be computed using only first derivative calculations. Thus, by ignoring the summation on the right hand side of (6.11), we obtain a quasi-Newton optimization scheme known as the *Gauss-Newton* algorithm. Combining this approximation to

(6.11) with (6.4), we note that the Gauss-Newton algorithm essentially chooses the search direction at each iteration as the least squares solution to the problem

$$\mathbf{J}\left(\mathbf{f}^{(n)}\right)\mathbf{d}^{(n)} = -\mathbf{e}\left(\mathbf{f}^{(n)}\right) \tag{6.12}$$

From Chapter 5, we know that if $\mathbf{J}$ is ill-conditioned, there will be difficulties in reliably solving (6.12). To avoid these problems, the *Levevberg-Marquardt* method employs a Tikhonov-type solution to (6.12)

$$\mathbf{d}^{(n)} = -\left[\mathbf{J}\left(\mathbf{f}^{(n)}\right)^T\mathbf{J}\left(\mathbf{f}^{(n)}\right) + \lambda\mathbf{I}\right]^{-1}\mathbf{J}\left(\mathbf{f}^{(n)}\right)^T\mathbf{e}\left(\mathbf{f}^{(n)}\right) \tag{6.13}$$

Methods for choosing λ are discussed in [70, Section 10.2] as are possibilities for approaching nonlinear least squares problems when the second term in (6.11) cannot be ignored.

6.2 Regularization II: Edge Preservation

As described in § 5.3, Tikhonov regularization methods are motivated by a desire to suppress high frequency, large amplitude artifacts in a reconstruction. Mathematically, this amounts to adding a term to the cost function which penalizes either the norm of the object itself or the norm of its gradient. Such *smoothness penalties* were shown to successfully remove the artifacts, but at the cost of blurring important object features such as edges or other area of rapid transition. For many problems, especially in imaging, these edges are of great practical importance for subsequent analysis of the resulting reconstructions. In particular, they may be needed to segment features of interest (such as tumors in a medical image or material flaws in nondestructive evaluation) from a nominal background. In many geophysical problems, the stratigraphy of the subsurface is characterized by edges between layers. Motivated by these and related concerns, there has been significant work done over the past 10-15 years in *edge-preserving regularization* which attempts to modify the Tikhonov approach so that edges are better recovered while still suppressing artifacts.

Let us start with the Tikhonov problem

$$\hat{\mathbf{f}} = \arg\min \|\mathbf{g} - \mathbf{K}\mathbf{f}\|_2^2 + \lambda\|\mathbf{L}\mathbf{f}\|_2^2 \tag{6.14}$$

The smoothness penalty is embodied in the second term where $\mathbf{L}$ in this context is taken to be an approximation to the gradient operator. To arrive at edge preserving methods, it is useful to rewrite the Tikhonov penalty as

$$\|\mathbf{L}\mathbf{f}\|_2^2 = \sum_k \left|[\mathbf{L}\mathbf{f}]_k\right|^2 = \sum_k \phi\left([\mathbf{L}\mathbf{f}]_k\right) \tag{6.15}$$

where $[\mathbf{L}\mathbf{f}]_k$ is the k-th element of the vector $\mathbf{L}\mathbf{f}$ and $\phi(x) = x^2$ in the Tikhonov case. In the statistical estimation literature, ϕ is known as a *potential function* and its first derivative is called the *influence function*. The smoothness inherent in Tikhonov reconstruction comes from the quadratic nature of ϕ which assumes that any large gradients in an object must be due to noise and hence are penalized rather aggressively. Unfortunately, image features such as edges also possess large

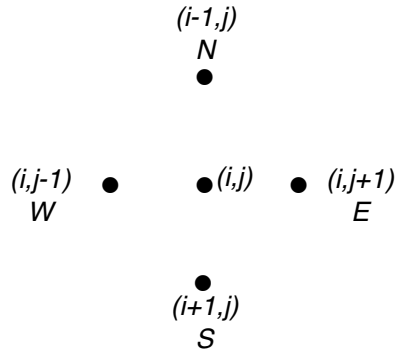


Figure 6.2: Indexing notation for pixel-based edge preserving regularization

gradients. By appropriately modifying ϕ however we can in fact recover these edges without adding noise.

Thus, following [13] we consider the following modified form of (6.14) as applied specifically to two dimensional problem

$$\hat{\mathbf{f}} = \arg\min_{\mathbf{f}} J(\mathbf{f}) \qquad J(\mathbf{f}) = \frac{1}{2}J_1(\mathbf{f}) + \frac{1}{2}\lambda J_2(\mathbf{f}) \tag{6.16}$$

$$J_1(\mathbf{f}) = \|\mathbf{g} - \mathbf{K}\mathbf{f}\|_2^2 \tag{6.17}$$

$$J_2(\mathbf{f}) = \sum_k \phi\left([\mathbf{D}_x\mathbf{f}]_k\right) + \sum_k \phi\left([\mathbf{D}_y\mathbf{f}]_k\right). \tag{6.18}$$

The matrices $\mathbf{D}_x$ and $\mathbf{D}_y$ implement first order difference approximations to the horizontal and vertical components of the gradient respectively. As shown in Fig. 6.2, at pixel i, j, these operators are:

$$[\mathbf{D}_x\mathbf{f}]_{i,j} = \mathbf{f}_{i,j+1} - \mathbf{f}_{i,j} \tag{6.19}$$

$$[\mathbf{D}_y\mathbf{f}]_{i,j} = \mathbf{f}_{i+1,j} - \mathbf{f}_{i,j} \tag{6.20}$$

where we have been (and will continue to be) a bit abusive of notation switching back and forth between k-based lexicographic indexing and i, j-based pixel indexing of the elements of $\mathbf{f}$.

As we know from § 6.1, a solution to (6.16) requires that the gradient of J vanish. The gradient of J_1 is

$$\nabla J_1(\mathbf{f}) = \mathbf{K}^T\mathbf{K}\mathbf{f} - \mathbf{K}^T\mathbf{g}. \tag{6.21}$$

The components of the gradient of J_2 are

$$\frac{\partial J_2(\mathbf{f})}{\partial \mathbf{f}_{i,j}} = -\left[\phi'(\mathbf{f}_{i,j+1} - \mathbf{f}_{i,j}) + \phi'(\mathbf{f}_{i,j} - \mathbf{f}_{i,j-1}) + \phi'(\mathbf{f}_{i+1,j} - \mathbf{f}_{i,j}) + \phi'(\mathbf{f}_{i,j} - \mathbf{f}_{i-1,j})\right] \tag{6.22}$$

where $\phi'(x)$ is the first derivative of ϕ with respect to x. Now assuming that $\phi'(x)/x \to M < \infty$

as $x \to 0$, we can rewrite (6.22) as[1]

$$\frac{\partial J_2(\mathbf{f})}{\partial \mathbf{f}_{i,j}} = -\left[\lambda_E \mathbf{f}_{i,j+1} + \lambda_W \mathbf{f}_{i,j-1} + \lambda_S \mathbf{f}_{i+1,j} + \lambda_N \mathbf{f}_{i-1,j} - \lambda_C \mathbf{f}_{i,j}\right] \tag{6.23}$$

where

$$\begin{aligned} \lambda_E &= \frac{\phi'(f_{i,j+1} - f_{i,j})}{2(f_{i,j+1} - f_{i,j})} \qquad \lambda_W = \frac{\phi'(f_{i,j} - f_{i,j-1})}{2(f_{i,j} - f_{i,j-1})} \\ \lambda_S &= \frac{\phi'(f_{i+1,j} - f_{i,j})}{2(f_{i+1,j} - f_{i,j})} \qquad \lambda_N = \frac{\phi'(f_{i,j} - f_{i-1,j})}{2(f_{i,j} - f_{i-1,j})} \\ \lambda_C &= \lambda_E + \lambda_W + \lambda_S + \lambda_N \end{aligned} \tag{6.24}$$

Gathering the components of (6.23) together gives the gradient of J_2 as

$$\frac{\partial J_2(\mathbf{f})}{\partial \mathbf{f}} = -\left[\boldsymbol{\nabla}^2(\mathbf{f})\right]\mathbf{f} \tag{6.25}$$

where each row of the matrix $\boldsymbol{\nabla}^2(\mathbf{f})$ implements the filtering operation of (6.23) on the appropriate collection of elements of $\mathbf{f}$.

We have used the somewhat cumbersome $\boldsymbol{\nabla}^2$ notation because of the link between (6.23) and a more traditional Laplacian filtering operation. In the image processing literature approximations to

$$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}$$

are routinely used for the detection of edges in images. Using the pixel indexing in Fig. 6.2, the standard discrete Laplacian is

$$\left[\boldsymbol{\nabla}^2 \mathbf{f}\right]_{i,j} = \mathbf{f}_{i,j+1} + \mathbf{f}_{i,j-1} + \mathbf{f}_{i+1,j} + \mathbf{f}_{i-1,j} - 4\mathbf{f}_{i,j} \tag{6.26}$$

which is identical to (6.23) for $\lambda_E = \lambda_W = \lambda_S = \lambda_N = 1$.[2] Moreover, it is not difficult to verify that if $\mathbf{L}$ in a Tikhonov regularization scheme is an approximation to the gradient then $\mathbf{L}^T\mathbf{L}$ is essentially $\boldsymbol{\nabla}^2$.[3] Thus, we can interpret (6.23) and (6.25) as an $\mathbf{f}$-dependent Laplacian filter. The issue we face now is the design of ϕ to achieve edge preservation.

More precisely we enumerate a collection of properties such a ϕ should possess. First, for regions of the image in which there are no manifest edges, ϕ should be chosen such that the resulting regularizer behaves like a standard Tikhonov; encouraging smoothness. If

$$\lim_{x \to 0} \frac{\phi'(x)}{2x} = M < \infty$$

then according to their definitions, for regions of an image where the intensity is slowly varying, λ_E, λ_W, λ_N, and λ_S will all go to M and (6.23) will go to (6.26) as desired. In this case, locally

[1] EXERCISE: Prove this

[2] EXERCISE: Verify this captures edges.

[3] EXERCISE: Again verify

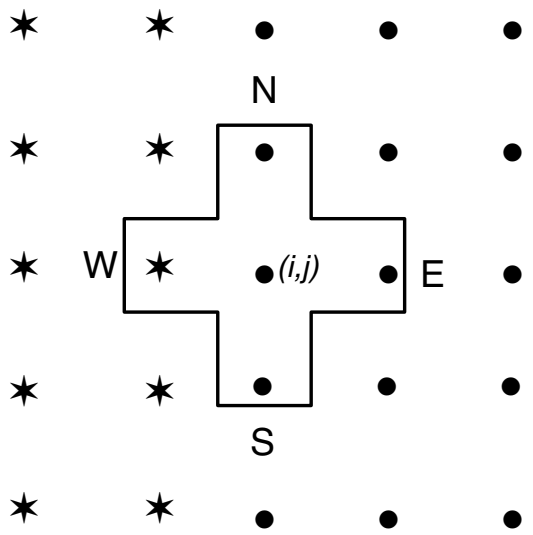


Figure 6.3: Preserving a vertical edge

at least, the edge preserving regularizer acts just like a Tikhonov regularizer. Next, as illustrated in Fig. 6.3, suppose that there is a discontinuity say between columns j and $j+1$. In this case all of the differences in (6.23) will be small except for the "westward" one. To preserve this edge in the reconstruction we want to ensure there is no penalty in the westward direction. This can be achieved if

$$\lim_{x\to\infty} \frac{\phi'(x)}{2x} = 0 \tag{6.27}$$

In such a case, (6.23) becomes

$$\frac{\partial J_2(\mathbf{f})}{\partial \mathbf{f}_{i,j}} \propto -\left[\mathbf{f}_{i,j+1} + \mathbf{f}_{i+1,j} + \mathbf{f}_{i-1,j} - 3\mathbf{f}_{i,j}\right]. \tag{6.28}$$

Hence again referring to Fig. 6.3 we will now encourage smoothness only over the region with pixels similar to that at location i, j. In other words if (6.27) holds, we would have a space-dependent Laplacian regularizer which performs no smoothing across discontinuities.

Formally, these ideas were assembled into a theory in [13] which required the following properties of ϕ for such a function to be edge-preserving:

1. $\phi(x) > 0$ and $\phi(0) = 0$
2. $\phi(x)$ should be even
3. $\phi(x)$ should be continuously differentiable
4. $\phi'(x)/2x$ should be continuous

5. $\phi'(x)/2x$ should be monotonically decreasing for $x \geq 0$ so that there is a one to one correspondence between the value of the gradient of $\mathbf{f}$ and the penalty it receives in terms of edge-preservation.

In fact, many such ϕ exist and, as recognized by the authors of [13], had been used in image processing for quite a while precisely to obtain sharper edge structure in a number of image restoration and denoising problems. A few of the more common are listed in Table 6.1.

$\phi(x)$	$\phi'(x)/2x$	Citation
$\frac{x^2}{1+x^2}$	$\frac{1}{(1+x^2)^2}$	[]
$\log(1+x^2)$	$\frac{1}{1+x^2}$	[]
$2\sqrt{1+x^2}-2$	$\frac{1}{\sqrt{1+x^2}}$	[]
$2\log[\cosh(x)]$	$\frac{\tanh(x)}{x}$ $x \neq 0$ and 1 for $x=0$	[]
$\sqrt{x^2+\beta^2}$ β small	$\frac{1}{\sqrt{\beta^2+x^2}}$	[]

Table 6.1: Edge Preserving Regularization Functions and Their Derivatives

A number of options exist for solving the edge preserving optimization problem embodied in (6.16)– (6.18). Gradient decent methods such as steepest decent and nonlinear conjugate gradient require only the derivative of the cost function with respect to the unknowns. This information can be assembled from equations (6.21)–(6.25) as well as Table 6.1 and is concisely summarized as

$$\nabla J(\mathbf{f}) = \left[\mathbf{K}^T\mathbf{K} + \lambda\boldsymbol{\nabla}(\hat{\mathbf{f}})\right]\mathbf{f} - \mathbf{K}^T\mathbf{g}. \tag{6.29}$$

Setting (6.29) equal to zero gives an implicit definition for the solution to the optimization problem

$$\left[\mathbf{K}^T\mathbf{K} + \lambda\boldsymbol{\nabla}(\hat{\mathbf{f}})\right]\mathbf{f} = \mathbf{K}^T\mathbf{g} \tag{6.30}$$

immediately suggesting an iterative scheme

$$\mathbf{f}^{(n+1)} = \left[\mathbf{K}^T\mathbf{K} + \lambda\boldsymbol{\nabla}(\mathbf{f}^{(n)})\right]^{-1}\mathbf{K}\mathbf{g} \tag{6.31}$$

which, when run to convergence gives $\hat{\mathbf{f}}$. It turns out that (6.31) is closely related to a Newton method for solving the problems. Specifically, it is not hard to show[4] that this iteration is equivalent to

$$\mathbf{f}^{(n+1)} = \mathbf{f}^{(n)} - \left[\mathbf{K}^T\mathbf{K} + \lambda\boldsymbol{\nabla}(\mathbf{f}^{(n)})\right]^{-1}\nabla J(\mathbf{f}^{(n)}) \tag{6.32}$$

where the matrix in brackets is an approximation to the Hessian of J.

Other issues to cover

- Link to total variation methods
- Half quadratic optimization methods

[4]EXERCISE

6.3 Nonlinear physical models

An second class of nonlinear inverse problems arises when the physics does not allow for a simple linear relationship between the data and the object. While there are a wide array of non-linear inverse problems depending on the physical process being modeled and the class of sensors being used, in this manuscript, we restrict our attention to one of the more widely studied of such problems: the non-linear inverse scattering problem for k^2 based on the Helmholtz model developed in Chapter 3. Even with this specialization, the literature remains quite extensive (see for example [18, Chapter 9], [49, Chapter 5], [69, Chapter 7]). Hence we shall constrain even further the scope of coverage to a presentation of issues related to the solution of these problems using the optimization framework of § 6.1.

As was mentioned at the end of § 6.2, making use of these algorithms requires the calculation of gradient (or Jacobian) information related to the cost function. As we discuss more formally shortly, such calculations ultimate require the derivative of the field, $\phi(\mathbf{r})$ with respect to changes in $k^2(\mathbf{r})$. Here we look at two methods for obtaining this sensitivity information. The first is based on a partial differential equation formulation of the underlying physics as in (3.10)while the second arises from the integral equation formulation in (3.25)– (3.28).

6.3.1 Adjoint Field Calculations

The adjoint field approach to gradient calculation starts with the PDE model for the scattering physics:

$$\nabla^2\phi(\mathbf{r}) + k^2(\mathbf{r})\phi(\mathbf{r}) = -s(\mathbf{r}) \tag{6.33}$$

plus appropriate boundary conditions. Let us assume for simplicity that the data we have for the inversion are sampled values of the field collected at point receivers[5]. We write the data-oriented part of the cost function involved in the inversion routine as

$$J_1\left[k^2(\mathbf{r})\right] = \frac{1}{2}\sum_m |\phi(\mathbf{r}_m, k^2(\mathbf{r})) - g(\mathbf{r}_m)|^2 \tag{6.34}$$

where $g(\mathbf{r}_m)$ is the datum collected at the receiver located at position $\mathbf{r}_m$ and $\phi(\mathbf{r}_m)$ is the solution to the PDE evaluated at $\mathbf{r}_m$ for a given $k^2(\mathbf{r})$.[6] As we know from § 6.1, the use of gradient based methods for solving the nonlinear problem require the derivative of the cost function with respect to the unknowns. In a framework where $k^2(\mathbf{r})$ is a function, the tradition notion of a derivative is replaced by a *functional*, or *Fréchet*, derivative of the cost function [72], [88, Chapter 2] [55, Chapter 7]. Informally, the Fréchet derivative is the linear operator that maps $\delta k^2(\mathbf{r})$ (a small perturbation in k^2,) into δJ_1 (the corresponding perturbation in the cost) and may be thought of as a functional generalization of a first order Taylor expansion in that

$$J_1\left[k^2(\mathbf{r}) + \delta k^2(\mathbf{r})\right] - J_1\left[k^2(\mathbf{r})\right] \equiv \delta J_1\left[k^2(\mathbf{r})\right] = \int \nabla J_1(\mathbf{r})\delta k^2(\mathbf{r})\, d\mathbf{r}. \tag{6.35}$$

[5]EXERCISE: Generalize to case where we have a linear functional of the fields

[6]To keep the notation from becoming too cluttered, we shall drop the explicit dependence of ϕ on k^2 in the following discussion

So the objective is to find an expression of the form (6.35) which relates a small change in k^2 to the change in the cost. The resulting structure of $\nabla J_1(\mathbf{r})$ is then the gradient function used in a steepest decent or conjugate gradient type of scheme for solving the nonlinear inverse scattering problem.

To begin, let us recall from complex variable theory that for any complex valued differentiable function, $f(x)$ where x is real-valued

$$\frac{d}{dx}|f(x)|^2 = 2\Re\left\{f^*\frac{df}{dx}\right\}. \tag{6.36}$$

from which we can conclude using (6.34) that

$$\delta J_1 = \Re\left\{\sum_m \left[\phi(\mathbf{r}_m, k^2(\mathbf{r})) - g(\mathbf{r}_m)\right]^* \delta\phi(\mathbf{r}_m)\right\} \tag{6.37}$$

where $\delta\phi(\mathbf{r}_m)$ is the perturbation in the field due to a small change in $k^2(\mathbf{r})$. Using (6.33) $\delta\phi(\mathbf{r}_m)$ may be found as the solution to the perturbed problem

$$\nabla^2\left(\phi + \delta\phi\right) + \left(k^2 + \delta k^2\right)\left(\phi + \delta\phi\right) = -s \tag{6.38}$$

plus boundary conditions. Under first order perturbation analysis we can ignore all terms in (6.38) which are quadratic (or higher) in terms of δ-quantities, such as $\delta\phi\delta k^2$. Thus we simplify (6.38) as

$$\nabla^2\delta\phi + \delta k^2\phi + k^2\delta\phi = 0 \tag{6.39}$$

where we have also used the fact that the unperturbed system satisfies $\nabla^2\phi + k^2\phi = -s$. Now we define the *adjoint field*, $\tilde{\phi}$ as the solution to

$$\nabla^2\tilde{\phi}(\mathbf{r}) + k^2\tilde{\phi}(\mathbf{r}) = -\tilde{s}(\mathbf{r}) \tag{6.40}$$

where the adjoint source is

$$\tilde{s}(\mathbf{r}) = \sum_m \left[g(\mathbf{r}_m) - \phi(\mathbf{r}_m)\right]^* \delta(\mathbf{r} - \mathbf{r}_m). \tag{6.41}$$

These definitions of both the adjoint source and the adjoint field allow us to rewrite (6.37) as

$$\delta J_1 = \Re\left\{\int \tilde{s}(\mathbf{r})\delta\phi(\mathbf{r})\,d\mathbf{r}\right\} \tag{6.42}$$

$$= -\Re\left\{\int \left[\nabla^2\tilde{\phi}(\mathbf{r}) + k^2\tilde{\phi}(\mathbf{r})\right]\delta\phi(\mathbf{r})\,d\mathbf{r}\right\} \tag{6.43}$$

According to (6.35), we would really like to see δk^2 under the integral rather than $\delta\phi$. This can be accomplished with the use of (6.38) and a little vector calculus as applied to the term $\int \delta\phi\nabla^2\tilde{\phi}\,d\mathbf{r}$. Specifically, using the identity[7]

$$\nabla\cdot\left[\tilde{\phi}\nabla\delta\phi\right] = \left(\nabla\tilde{\phi}\right)\cdot\left(\nabla\delta\phi\right) + \tilde{\phi}\nabla^2\delta\phi$$

[7]EXERCISE: Prove the identity

we have

$$\nabla \cdot \left[\tilde{\phi} \nabla \delta\phi - \delta\phi \nabla \tilde{\phi} \right] = \delta\phi \nabla^2 \tilde{\phi} + \tilde{\phi} \nabla^2 \delta\phi$$

so that

$$\int \delta\phi \nabla^2 \tilde{\phi} \, d\mathbf{r} = \int d\mathbf{r} \left[-\tilde{\phi} \nabla^2 \delta\phi + \nabla \cdot \left(\tilde{\phi} \nabla \delta\phi - \delta\phi \nabla \tilde{\phi} \right) \right]. \tag{6.44}$$

The $\nabla\cdot$ term in (6.44) can be shown to be equal to zero by the following argument:

1. Convert the volume integral to a surface integral using the divergence theorem.
2. Let the surface over which the integral is computed be a ball encompassing the perturbation δk^2 and allow the radius of the ball to go to infinity.
3. Assuming the fields obey the Sommerfeld radiation condition, (3.14), the resulting integral can be shown to go to zero [72].

After some algebra, using (6.44) with the second term on the right hand side equal to zero in (6.43) results in

$$\delta J_1 = -\Re \left\{ \int \tilde{\phi} \left[\nabla^2 \delta\phi + k^2 \delta\phi \right] \, d\mathbf{r} \right\} \tag{6.45}$$

but according to (6.39), the term in brackets in (6.45) is $-\delta k^2 \phi$. Hence

$$\delta J_1 = -\Re \left\{ \int \tilde{\phi}(\mathbf{r}) \phi(\mathbf{r}) \delta k^2(\mathbf{r}) \, d\mathbf{r} \right\} \tag{6.46}$$

which, on comparison to (6.35) shows that the Fréchet derivative for this problem must be

$$\nabla J_1(\mathbf{r}) = \Re \left\{ \tilde{\phi}(\mathbf{r}) \phi(\mathbf{r}) \right\}. \tag{6.47}$$

Thus, to find the gradient information for the inverse scattering problem in which there are N_s sources and N_r receivers requires a total of $N_s + N_r$ forward solves using the adjoint field approach per iteration of the underlying optimization method. To be more specific, the fields $\phi(\mathbf{r})$ satisfying (6.33) are required to compute the data residuals in (6.34). For N_s source function, N_s solves will be required. According to (6.40) and (6.41), adjoint fields are required at each receiver location as well. While the adjoint field for a given source requires the data residuals for that source; by the linearity of the adjoint problem (6.40), we can synthesize $\tilde{\phi}(\mathbf{r})$ as

$$\tilde{\phi}(\mathbf{r}) = \sum_m \left[g(\mathbf{r}_m) - \phi(\mathbf{r}_m) \right]^* \tilde{\phi}_m(\mathbf{r}) \tag{6.48}$$

where $\tilde{\phi}_m(\mathbf{r})$ is the solution to

$$\nabla^2 \tilde{\phi}_m(\mathbf{r}) + k^2 \tilde{\phi}_m(\mathbf{r}) = -\delta(\mathbf{r} - \mathbf{r}_m). \tag{6.49}$$

Hence at any iteration of the algorithm, N_s forward solves and N_r adjoint solves can be used to compute the required sensitivity information.

It is interesting to compare (6.47) with the Born approximation, (3.30). The kernel for the Born model is of the form $g(\mathbf{r}_m, r)\phi_b(\mathbf{r})$. The factor ϕ_b in the Born model is the "background field" which can be interpreted as the field produced by some source for a known k^2. The factor $g(\mathbf{r}_m, r)$ is the field arising from a δ source placed at the location of a receiver again for some nominal k^2. In the context of solving the nonlinear inverse scattering problem, at iteration n, this "nominal" k^2 is $k^{2,(n)}$. Hence we can identify ϕ in (6.47) with ϕ_b in the Born model and $\tilde{\phi}_m$ with g. So, the manner in which the gradient information required to solve the nonlinear inverse scattering problem is computed amounts to a succession of Born linearizations about the iterates produced as the algorithm proceeds.[8]

In addition to their use for gradient decent optimization methods, adjoint field approaches can also be used for Jacobian calculations required in Gauss-Newton or Levenburg-Marquardt algorithms. Examining (6.12) and (6.13), we see that search directions for these methods requires both the residual $\phi(\mathbf{r}_m, k^2(\mathbf{r})) - g(\mathbf{r}_m)$ and the Jacobian of the residual with respect to k^2. To adapt the adjoint field method to the finite dimensional case, we assume for simplicity that $k^2(\mathbf{r})$ is expanded in a pixel-type basis. In this case for a given source, the derivative of the m-th measurement with respect to the n-th pixel requires only the integral of $\tilde{\phi}_m(\mathbf{r})\phi(\mathbf{r})$ over the support of the m-th pixel. Note that some care must be taken here as the adjoint field method was defined for complex-valued residuals while the algorithms in § 6.1.2 assumed real valued residuals.

6.3.2 Integral Equation Method

Using a discretized form of the integral equation formulation of the scattering problem developed in § 3.3.3, it is possible to compute the *exact* derivative of the cost function with respect to each pixel value of interest. The resulting expression also lend some light onto the adjoint field methods for sensitivity calculation.

From § 3.3.3, the physical model of interest here is summarized by a pair of coupled integral equations

$$\phi(\mathbf{r}) + \int g(\mathbf{r}, \mathbf{r}')k_s^2(\mathbf{r}')\phi(\mathbf{r}')\, d\mathbf{r}' = \phi_b(\mathbf{r}) \tag{6.50}$$

$$\phi_s(\mathbf{r}_m) = \int g(\mathbf{r}_m, \mathbf{r}')\phi(\mathbf{r}')k_s^2(\mathbf{r}')\, d\mathbf{r}' \tag{6.51}$$

where we have decomposed $k^2(\mathbf{r})$ into background and perturbation (or scattering) components. The background field ϕ_b and the Green's function, g, are both computed for the background and the objective of the problem is to recover the perturbation, k_s^2, given $\phi_s(\mathbf{r}_m)$, scattered fields collected at receiver locations $\mathbf{r}_m$. For simplicity, say we discretize (6.50) and (6.51) using the method of moments with flat-top basis functions as in the top of Fig. 3.6. We then arrive at the matrix-vector

[8] EXERCISE IDEAS:

1. Functional derivative for D
2. What if the forward model *is* linear. Show the Fréchet derivative is what it should be.
3. The addition of a regularizer to the problem

equations

$$(\mathbf{I} + \mathbf{G}\mathcal{D}(\mathbf{f}))\,\boldsymbol{\phi} = \boldsymbol{\phi}_b \tag{6.52}$$

$$\mathbf{g} = \mathbf{G}_m\mathcal{D}(\boldsymbol{\phi})\mathbf{f} \tag{6.53}$$

where

- $\mathbf{f}$ is the vector of unknown pixel values for $k_s^2(\mathbf{r})$
- $\boldsymbol{\phi}$ is the vector of pixel value for $\phi(\mathbf{r})$
- $\mathbf{G}$ and $\mathbf{G}_m$ hold the discretized Green's functions in (6.50) and (6.51) respectively
- $\mathbf{g}$ is he vector of observed scattered fields
- $\mathcal{D}(\mathbf{x})$ is the diagonal matrix formed from the elements of the vector $\mathbf{x}$.

Solving (6.52) for $\boldsymbol{\phi}$ and using the result in (6.53) gives the discrete analog to (3.28)

$$\mathbf{g} = \mathbf{G}_m\mathcal{D}\left\{(\mathbf{I} + \mathbf{G}\mathcal{D}(\mathbf{f}))^{-1}\,\boldsymbol{\phi}_b\right\} \tag{6.54}$$

Recalling the discussion in § 6.1.2, the Jacobian required in for a Gauss-Newton or Leveberg-Marquardt type approach is the matrix whose i,j-th component is the derivative of g_i with respect to f_j. Denoting this matrix as $\partial\mathbf{g}/\partial\mathbf{f}$ and using tedious but straightforward linear algebra we have

$$\frac{\partial\mathbf{g}}{\partial\mathbf{f}} = \mathbf{G}_m\mathcal{D}(\boldsymbol{\phi}) + \mathbf{G}_m\mathcal{D}(\mathbf{f})\frac{\partial\boldsymbol{\phi}}{\partial\mathbf{f}} \tag{6.55}$$

where $\boldsymbol{\phi} = (\mathbf{I} + \mathbf{G}\mathcal{D}(\mathbf{f}))^{-1}\,\boldsymbol{\phi}_b$. We compute the Jacobian of the $\boldsymbol{\phi}$ with respect to $\mathbf{f}$ one column at a time. Differentiating both sides of (6.52) with respect to $\mathbf{f}_i$ we have

$$[\mathbf{I} + \mathbf{G}\mathcal{D}(\mathbf{f})]\,\frac{\partial\boldsymbol{\phi}}{\partial\mathbf{f}_i} = -\mathbf{G}\left[\frac{\partial}{\partial\mathbf{f}_i}\mathcal{D}(\mathbf{f})\right]\boldsymbol{\phi} \tag{6.56}$$

Because the partial derivative of $\mathcal{D}(\mathbf{f})$ with respect to $\mathbf{f}_i$ is a matrix with all zeroes except for a single 1 in the i,i location we can simplify (6.56) to

$$\frac{\partial\boldsymbol{\phi}}{\partial\mathbf{f}_i} = -\left[\mathbf{I} + \mathbf{G}\mathcal{D}(\mathbf{f})\right]^{-1}\mathbf{G}\mathcal{D}(\boldsymbol{\phi})\mathbf{e}_i \tag{6.57}$$

where $\mathbf{e}_i$ is the i-th unit vector. Using (6.57), gives

$$\frac{\partial\boldsymbol{\phi}}{\partial\mathbf{f}} = \left[\frac{\partial\boldsymbol{\phi}}{\partial\mathbf{f}_1} \mid \frac{\partial\boldsymbol{\phi}}{\partial\mathbf{f}_2} \mid \frac{\partial\boldsymbol{\phi}}{\partial\mathbf{f}_3} \mid \cdots\right] = \left[\mathbf{I} + \mathbf{G}\mathcal{D}(\mathbf{f})\right]^{-1}\mathbf{G}\mathcal{D}(\boldsymbol{\phi}). \tag{6.58}$$

Finally, by substituting (6.58) into (6.56) we arrive at

$$\frac{\partial\mathbf{g}}{\partial\mathbf{f}} = \mathbf{G}_m\mathcal{D}(\boldsymbol{\phi}) + \mathbf{G}_m\mathcal{D}(\mathbf{f})\left[\mathbf{I} + \mathbf{G}\mathcal{D}(\mathbf{f})\right]^{-1}\mathbf{G}\mathcal{D}(\boldsymbol{\phi}). \tag{6.59}$$

Note that by dropping the second term in (6.58), we have an expression quite closely related to that obtained using the adjoint field approach at the end of the last section. The first term on the right hand side of (6.58) looks very much like a discretized form of an integral operator whose kernel is $g(\mathbf{r}_m, \mathbf{r})\phi(\mathbf{r})$. This is identical to the adjoint approach except for the fact that $g(\mathbf{r}_m, \mathbf{r})$ here is the Green's function computed about the nominal background k_b^2 which will not change as the optimization method proceeds. This is in contrast to the adjoint approach where g is replaced by $\tilde{\phi}_m$, the adjoint field computed for the current iterate of k^2.

- The full nonlinear inverse scattering problem revisited
- Sensitivity calculations methods: adjoint field, integral equation based
- Contrast source inversion method

6.4 Geometric Inverse Methods

- Inverting for parametric models (spheres, splines, etc).
- Level set inverse methods

6.5 Exercises

6.1 Recall that the edge preserving linear inverse problem is to recover an estimate of the object f in a way which solves the following no-quadratic optimization problem:

$$\hat{f} = \arg\min_f \|g - Kf\|_2^2 + \lambda^2 \Omega(Df) \tag{6.60}$$

where Ω is an edge preserving functional and Df represents the gradient of the object.

For the remainder of this problem, let $N = 128$ and construct K as follows in Matlab:

```
N    = 128;
n = linspace(0,1,N);
dx = n(2)-n(1);
hsig = .0236;
K =zeros(N);
for i = 1:N
   for j = 1:N
      K(i,j) = dx/(hsig*sqrt(2*pi))*...
               exp(-((i-j)*dx)^2 / (2*hsig*hsig));
   end
end
```

In later parts of this problem, we will look at the performance of these methods for an f constructed as follows:

```
f = zeros(N,1);
for idx = 1:N
  if ((idx > 20) & (idx < 45))
    f(idx) = 1;
  end
  tmp = 3*exp(-( ((idx-90)^2/80)));
  f(idx) = f(idx) + tmp;
```

```
end
```

Assuming that f exists in the Matlab workspace, and the variable SNR has been set to a desired noise level in dB, we generate noisy data according to the following:

```
g_clean = K*f1;
noise_var = norm(g_clean)/length(g_clean)*10^(-SNR/10);
g = g_clean + sqrt(noise_var)*randn(length(g_clean),1);
```

1. It is never really clear how to choose the regularization parameter for this problem. Assuming we use a Tikhonov rather than edge preserving (i.e. $\Omega(DF) = \|Df\|_2^2$, for each of the f's and at an SNR of 30 dB find the two parameters which minimize the difference between the true f and the one generated by the Tikhonov method as a function of the parameter. Comment on the stability of the "best" regularization parameter for each f. Comment on why this approach is not feasible in practice.
2. In Chapter 8 of the Vogel text, there is a discussion of three methods for solving the optimization problems associated with Edge Preserving Regularization. It turns out that the one which converges fastest is the primal-dual approach of Section 8.2.5. While the theory of this method is a bit more than we want to cover in this class, a pseudo-code implementation of the method is provided on page 141 for the 2D problem. Using this code as a base, write a simplified version of the algorithms for the 1D problem. For this problem, use the ϕ function indicated in equation (8.38) of the text.
3. For each of the two f's try the edge preserving method. Set the regularization parameter equal to the one found previously in this problem. Comment on how these results differ from that of the best Tikhonov output.
4. How do the results change if we lower the SNR to 10? What about increasing the value of `hsig` used to build K to 0.1.

6.2 In many problems, the physics dictates that the quantity being recovered must be bounded from below typically by zero. For example, inverse problems associated with physics require that the sound speed be estimated. Clearly, this quantity must be non-negative. Similarly, electrical conductivity is also required to be non-negative. There are a number of ways of incorporating such bounds into the reconstruction, some more rigorous than others. Here we want to look at one: a change of variable. To see how this works, consider the typical linear inverse problem:

$$\hat{f} = \arg\min_f \|g - Af\|_2^2 + \lambda^2 \|Df\|_2^2 \tag{6.61}$$

with D the first difference matrix. Now we want to add the condition that each element of f is greater than or equal to zero.

1. Why can't we use the normal least squares type of solution still?

2. Now let us suppose that instead of estimating the elements of f, we assume that f can be associated with a second vector, h via $f_i = h_i^2$. In Matlab notation, `f = h.* h`. Now the problem is to find $\hat{h}$ according to

$$\hat{h} = \arg\min_f \|g - A(h. * h)\|_2^2 + \lambda^2 \|D(h. * h)\|_2^2 \tag{6.62}$$

and define $\hat{f} = \hat{h}. * \hat{h}$. Develop a steepest decent algorithm for solving this problem.

3. Implement and test the algorithm using the following code for K and f:

```
% $$$ DEFINITION OF THE MATRIX K.
N    = 128;
n = linspace(0,1,N);
dx = n(2)-n(1);
hsig = .05;
K =zeros(N);
for i = 1:N
   for j = 1:N
      K(i,j) = ...
          dx/(hsig*sqrt(2*pi))*...
                   exp(-  ( (i-j)*dx)^2 / (2*hsig*hsig)) ...
          - dx/(hsig*sqrt(2*pi))*...
                   exp(-  ( (i-j-N/4)*dx)^2 / (2*hsig*hsig));
   end
end

% $$$ DEFINITION OF THE FUNCTION F.
f = zeros(N,1);
for idx = 1:N
  if ((idx > 20) & (idx < 45))
    f(idx) = 1;
  end
  tmp = 3*exp(-( ((idx-90)^2/80)));
  f(idx) = f(idx) + tmp;
end
```

The regularization matrix should be a discrete derivative and the optimal regularization parameter should be selected like in the last problem.The SNR should be 10. Please compare results to the unconstrained Tikhonov method.

Appendix A

A Brief Review of Probability

Probability is a mathematical formalism for quantifying ideas concerning outcomes of experiments that are (or appear to be or are well approximated as) random. The study of this topic is extensive covered in a range of texts and research monographs. Moreover, there exists interesting connections between many of the inverse methods developed in this book and techniques drawn from the area of probabilistic inference. Here we content ourselves with a brief overview of a few relevant concepts and point the reader to references such as [22, 26, 88] for further, more detailed analysis.

A.1 Basic Concepts

The fundamental components of a probabilistic model are:

1. The *sample space*, $\mathcal{S}$, defined to be "The finest-grain, mutually exclusive, collectively exhaustive listing of all possible outcomes of a model of an experiment" [26, Section 1.2].

2. *Events*: possible outcomes of an experiment which may not be finest-grain, mutually exclusive, or collectively exhaustive.

3. A *probability measure* which is a way of assigning to an event a number between zero and one indicating the odds of the event actually occurring as an outcome to an experiment.

As a simple example of the above, consider an experiment of rolling a pair of six sided dice. The sample space is the collection of all thirty six possible pairs of spots that may be observed as a result of the role. In this simple example, event constitute subsets of these thirty six outcomes. A simple event is "Die one shows a six and die two shows a three" which is just a single element of the sample space. More complex event can also be considered such as "The sum of the two numbers showing on the dice is even" or "Die one shows a four." If the dice are fair (no one number more likely than any other), the measure for this experiment assigned the probability of 1/36 to each of the thirty six possible outcomes. The probability of a particular event then is the sum of the probabilities for the constituent elements in the sample space. Letting x represent the event we denote by $P(x)$ the probability assigned to x. Thus if x is comprised of the union of elements x_i in

the sample space we have

$$P(x) = \sum_i P(x_i) \tag{A.1}$$

A key issue in the area of probabilistic modeling is the incorporation of known information into probability calculations. In the context of inverse problems this arises naturally enough via data. If we observe a single piece of random data, y which we know carries information about a quantity of interest, x, it is natural to compare the probability that x has occurred knowing the outcome y to the probability that x occurs in the absence of this information. The tool for exploring this issue is the *conditional probability* of x given y, $P(x|y)$. To determine $P(x|y)$, we look at that subset of elements of the sample space where y is known to have occurred. Call this set Y. The conditional probability of x is then the sum over all elements of Y where x is also true. To ensure that the probability is normalized to one, we must scale this result by $P(y)$. Thus mathematically we have

$$P(x|y) = \frac{1}{P(y)} \sum_{i \in Y} P(x_i). \tag{A.2}$$

The sum in (A.2) is nothing more than the probability that both x and y are true which we write as $P(xy)$. Hence

$$P(x|y) = \frac{P(xy)}{P(y)}. \tag{A.3}$$

When knowing y tells us nothing about the probability of x, the two events are called *independent*. In this case, $P(x|y) = P(x)$, the original probability of x and we have the important relationship that for independent events x and y, $P(xy) = P(x)P(y)$.

Suppose that we can write $\mathcal{S}$ as the union of mutually exclusive events x_i; *i.e.* $\mathcal{S} = \cup_i x_i$ and for $i \neq j$ $x_i \cap x_j = \emptyset$. Then the probability that an event y occurs is equal to [75, Section 2.3]

$$P(y) = \sum_i P(y|x_i)P(x_i). \tag{A.4}$$

From the definition of conditional probability P(xy) = P(x—y)P(y) = P(y—x)P(x)

$$P(x_i|y) = \frac{P(y|x_i)P(x_i)}{P(y)} = \frac{P(y|x_i)P(x_i)}{\sum_i P(y|x_i)P(x_i)} \tag{A.5}$$

which is known as *Bayes theorem*.

A.2 Random Variable

For cases where a numerical value can be assigned to an outcome of an experiment, one can build on the results of the previous section though the notion of a *random variable*. Formally a random variable is a function that assigns to a point in the sample space a probability. For example we may have x as a random variable which represents the number of spots seen on the role of a single fair die. The function that assigns $1/6$ to each of the numbers $1, 2, 3, 4, 5, 6$ in the sample space is called the *probability mass function* (PMF) and denoted $p(x)$.

Experiments are not restricted to classes of event which assume a finite (or even countable) number of outcomes. If one thinks of the time between successive trains on the subway as a random variable, clearly this quantity can take on any value between 0 and ∞. The appropriate generalization of the probability mass function for continuous random variables is the *probability density function* (PDF), $f(x)$. Using the PDF, the probability of an event as the integral over all possible outcomes of $f(x)$. If $f(t)$ is the PDF for the waiting time (in minutes) for the next train then the probability of the event $A =$ "We will have to wait between 10 and 12 minutes" is

$$P(A) = \int_{10}^{12} f(t)dt.$$

As another example, assuming that $f(x)$ is continuous (which will always be the case for us here) the probability that a continuous random variable assumes a value between x_0 and $x_0 + \delta x$ for an infinitesimal δx is

$$P(x_0 \leq x \leq x_0 + \delta x) = \int_{x_0}^{x_0+\delta x} f(x)dx \approx f(x_0)\delta x.$$

This results indicates that $f(x)$ is closely related to an actual probability. From this it follows that for a function to be a PDF, it must be positive and its integral over all x must equal 1.

Closely associated with random variables is the notion of *expectation*. Say that we observe not x, but a function of this quantity $y = g(x)$. Since x is random, so too will be $y = g(x)$. While there exist well defined methods for deriving the distribution $f(y)$ [26, Section 2-14], a simpler quantity to obtain is the *expected value* of y, $E[y] = E[g(x)]$ defined as

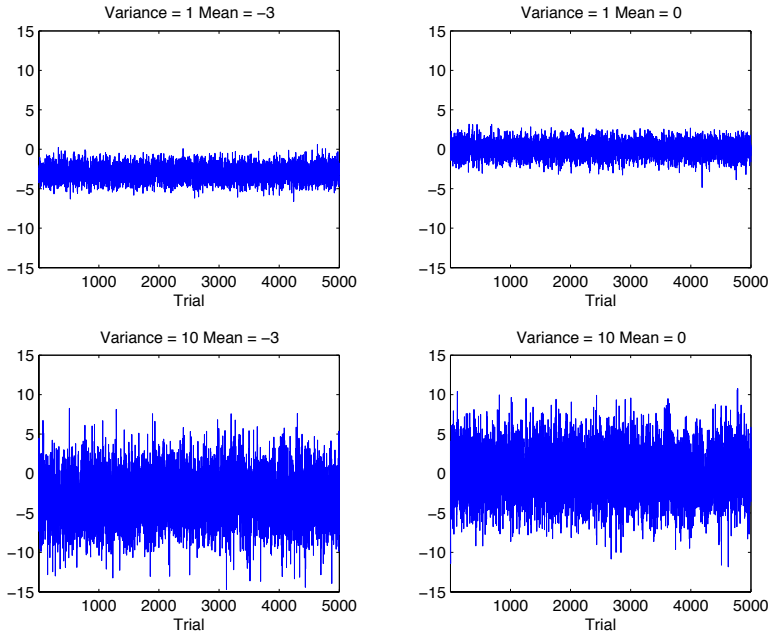
$$E[y] = \int g(x)f(x)dx \tag{A.6}$$

Basically, one would expect that if we were to measure y many, many times with each measurement independent of all the others then average value we see should be influenced in proportion to the quantity of probability associated with the underlying x. Two important expectations for us here are the expected value (or mean), $E[x]$ and the variance $E[(x - E[x])^2]$. Given no other information, if one had to guess the result of an experiment involving x, the number produced would be $E[x]$. The variance measures the expected spread of the true value of x about $E[x]$. The larger the variance, the more likely that the outcome x will be far from $E[x]$. As the variance decreases, one expects to see a tighter spread of values clustered more closely to $E[x]$.

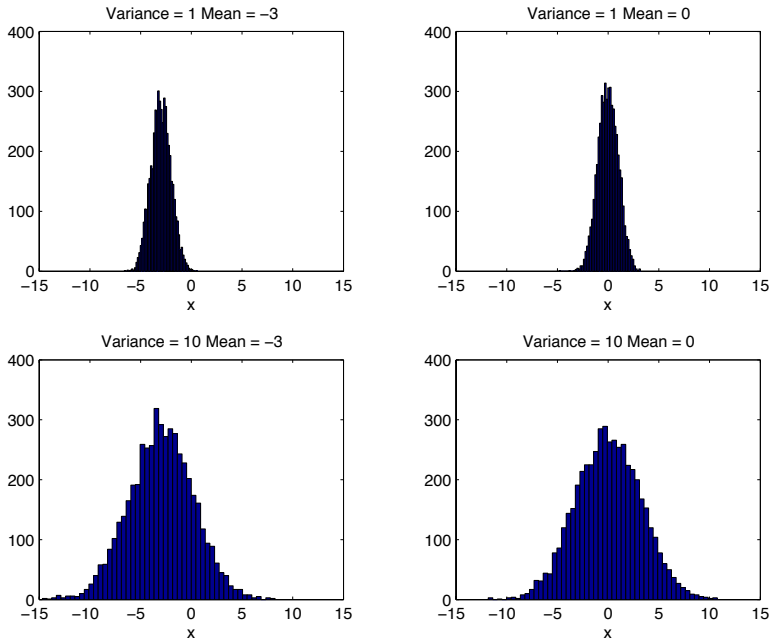
To make these ideas a bit more concrete, let us take a canonical example. While there exist a variety of analytically specified PDFs that are commonplace in mathematical modeling and engineering, by far the most commonly used is the *normal* or *Gaussian* probability density function defined as

$$f(x) = \frac{1}{\sqrt{2\pi\nu^2}} e^{-(x-\bar{x})^2/2\nu^2}. \tag{A.7}$$

it is not hard to show that for the normal density, the mean is $\bar{x}$ and the variance is ν^2. In Fig. A.1(a), we plot collections of 5000 independently generated samples from a normal random variable as we change the mean and variance. Histograms of the results are displayed in Fig. A.1(b). The height of each bar in the histograms represents the number of times in each set of 5000, the



(a) Independent samples of a normal random variable for different means and variances



(b) Histograms for the corresponding plots in (a)

Figure A.1: Samples and associated histograms for normally distributed random variables.

value of the random variable fell within the region spanned by the width of the bar. If we were to normalize these histograms by 5000, they would represent an approximation to the corresponding PDF. These results in Fig. A.1 show that as $\bar{x}$ is changes the plots in (a) tend to cluster about the appropriate values while those in (b) shift along the abscissa to be centered on the value of $\bar{x}$. Increasing the variance, ν^2 is also seen to increase the spread of the values seen for a given set of 5000 samples. Thus the histograms in (b) are more spread out while the plots in (a) show wider variation about the mean value.

A.3 Jointly Distributed Random Variables

Given more than one random variable, it is both possible and quite useful to define a joint density function (or mass function in the discrete case). In the case of N random variables $x_1, x_2, \ldots x_N$, we write the PDF as $f(x_1, x_2, \ldots, x_N)$ or using vector notation $f(\mathbf{x})$ where $\mathbf{x}^T = [x_1\, x_2\, \ldots\, x_N]$. As in the univariate case, the joint PDF must be both positive and integrate to unity over the whole sample space. Moreover, $f(\mathbf{x})$ evaluated at some $\mathbf{x}_0$ can be interpreted as the probability of the event $A =$ "$x_1 \in [x_{1,0}, x_{1,0} + \delta x_1]$, and $x_2 \in [x_{2,0}, x_{2,0} + \delta x_2]$, and $\ldots, x_N \in [x_{N,0}, x_{N,0} + \delta x_N]$" which is

$$P(A) = f(\mathbf{x}_0)\delta x_1 \delta x_2 \cdots \delta x_N$$

Defined in this manner, it is possible to extend naturally the notion of conditional probability to continuous random variables. For example, as explained in [26, Section 2-12], the PDF of the random variable x conditioned on knowledge of a second random variable y is

$$f(x|y) = \frac{f(x,y)}{f(y)}.$$

The idea can obviously be extended to cases where there are more than two random variables.

Multivariate expectations and conditional expectations are defined in the natural manner as

$$E[g(\mathbf{x})] = \int g(\mathbf{x}) f(\mathbf{x}) d\mathbf{x} = \int g(x_1, x_2, \ldots, x_N) f(x_1, x_2, \ldots, x_N) dx_1 dx_2 \ldots dx_N$$

$$E[g(\mathbf{x})|\mathbf{y}] = \int g(\mathbf{x}) f(\mathbf{x}|\mathbf{y}) d\mathbf{x}.$$

The generalization of the mean of a scalar random variable is the mean vector $E[\mathbf{x}] = \bar{\mathbf{x}}$ of multivariate density. The ith element of $E[\mathbf{x}]$ is just $E[x_i]$. The multivariate extension of the variance is a bit more involved. Now, we can consider the co-variation of each x_i with respect to the other x_j. This leads to the notion of the covariance matrix, $\mathbf{Q}$, associated with the random vector $\mathbf{x}$

$$\mathbf{Q} = E[(\mathbf{x} - \bar{\mathbf{x}})(\mathbf{x} - \bar{\mathbf{x}})^T].$$

Thus, $\mathbf{Q}$ is a symmetric matrix (actually, symmetric positive semi-definite [75, Section 8-1]) whose (m,n)-th entry is $E[(x_m - \bar{x}_m)(x_n - \bar{x}_n)]$. Note that the diagonal elements of $\mathbf{Q}$ are the variances of the individual components of $\mathbf{x}$. If the covariance matrix possess a Toeplitz structure so that $\mathbf{Q}_{i,j} = \mathbf{Q}_{i-j}$ then the random vector is said to be *stationary*.

The multivariate Gaussian density is

$$f(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^N |\mathbf{Q}|}} e^{(\mathbf{x}-\bar{\mathbf{x}})^T \mathbf{Q}^{-1}(\mathbf{x}-\bar{\mathbf{x}})/2} \tag{A.8}$$

where the mean of $\mathbf{x}$ is in fact given by $\bar{\mathbf{x}}$, the covariance matrix is $\mathbf{Q}$, and $|\mathbf{Q}|$ is the determinant of the matrix $\mathbf{Q}$[1]. As a shorthand, when $\mathbf{x}$ follows a distribution as in (A.8), we write $\mathbf{x} \sim N(\bar{\mathbf{x}}, \mathbf{Q})$. The multivariate Gaussian is used quite extensively (and sometimes even correctly) as a model for additive sensor noise in general and for inverse problems in particular. In many of these cases, the phrase "white" noise is employed as an indication that the samples are uncorrelated. This in turn gives rise to a diagonal covariance matrix $\mathbf{Q} = \text{diag}(\nu_1^2, \nu_2^2, \ldots, \nu_N^2)$. If in addition, the vector is stationary then all of the variances are the same and equal to ν^2 and $\mathbf{Q} = \nu^2 \mathbf{I}$.

[1]Here we consider only cases where $\mathbf{Q}$ is invertible and hence positive definite.

Bibliography

[1] Simon Arridge. Optical tomography in medical imaging. *Inverse Problems*, 15:R41–R43, 1999.

[2] Harry Bateman. *Higher Transcendental Functions*, volume II. Robert E. Krieger Publishing Co., Inc., 1981.

[3] Murat Belge, Misha Kilmer, and Eric Miller. Wavelet domain image restoration with adaptive edge-preserbing regularization. *IEEE Trans. Image Processing*, 9(4):597–608, April 2000.

[4] Murat Belge, Misha Kilmer, and Eric L. Miller. Efficient determination of multiple regularization parameters in a generalized L-curve framework. *Inverse Problems*, 18:1161–1183, 2002.

[5] M. Bertero. Linear inverse and ill-posed problems. In Hawkes Peter W, editor, *Advances in Electronics and Electron Physics*, volume 75, pages 1–120. Academic Press, Boston, 1989.

[6] M. Bertero, C. De Mol, and E. R. Pike. Linear inverse problems with discrete data. I: General formulation and singular system analysis. *Inverse Problems*, 1:301–330, 1985.

[7] M. Bertero, C. De Mol, and E. R. Pike. Linear inverse problems with discrete data, II, Stability and regularisation. *Inverse Probl.*, 4:573–594, 1988.

[8] N. Bleistein, J. K. Cohen, and J. W. Stockwell Jr. *Mathematics of Multidimensional Seismic Imaging, Migration and Inversion*, volume 13 of *Interdisciplinary Applied Mathematics*. Springer Verlag, 2001.

[9] D. A. Boas. A fundamental limitation of linearized algorithms for diffuse optical tomography. *Optics Express*, 1(13), 22 December 1997.

[10] Max Born and Emil Wolf. *Principles of Optics: Electromagnetic Theory of Propagation, Interference and Diffraction of Light*. Cambridge UNiversity Press, seventh edition, 1999.

[11] J. R. Bowler, S. J. Norton, and D. J. Harrison. Eddy current interaction with an ideal crack. II The inverse problems. *J. Applied Physics*, 75:8138–8144, 1994.

[12] Ronald N. Bracewell. *Two-Dimensional Imaging*. Prentice Hall, 1995.

[13] P. Charbonnier, L Blanc-Feraud, G. Aubert, and M. Barlund. Deterministic edge-preserving regularization in computed imaging. *IEEE Trans. Image Processing*, 6(2):298–311, February 1997.

[14] M. Cheney, D. Isaacson, and J.C. Newell. Electrical impedance tomogrpahy. *SIAM Review*, 40:85–101, 1999.

[15] Margaret Cheney. A mathematical tutorial on synthetic aperture radar. *SIAM Review*, 43:301–312, 2001.

[16] Margaret Cheney and Brett Borden. Microlocal structure of inverse synthetic aperture radar data. *Inverse Problems*, 19:173–194, 2003.

[17] W. C. Chew, G. P. Otto, W. H. Weedon, J. H. Lin, C. C. Lu, Y. M. Wang, and M. Moghaddam. Nonlinear diffraction tomography: T use of inverse scattering for imaging. *International Journal of Imagaing Systems and Technology*, 7:16–24, 1996.

[18] Weng Cho Chew. *Waves and Fields in Inhomogeneous Media*. Van Nostrand Reinhold, New York, 1990.

[19] D. Colton and P. Monk. A linear sampling method for the detection of leukemia using microwaves. *SIAM J. Applied Math.*, 58:926–941, 1998.

[20] David Colton and Peter Monk. A modified dual space method for solving the electromagnetic inverse scattering problem for an infinite sylinder. *Inverse Problems*, 10:87–107, 1994.

[21] L. M. Delves and J. L. Mohamed. *Computational Methods for Integral Equations*. Cambridge University Press, 1988.

[22] Guy Demoment. Image reconstruction and restoration: Overview of common estimation structures and problems. *IEEE Trans on Image Processing*, 37(12):2024–2036, 1989.

[23] A. J. Devaney. A filtered backprojection algorithm for diffraction tomography. *Ultrasonic Imaging*, 4:336–350, 1982.

[24] A. J. Devaney. Geophysical diffraction tomography. *IEEE Trans. on Geoscience and Remote Sensing*, GE-22(1):3–13, January 1984.

[25] Oliver Dorn, Eric L. Miller, and Carey Rapaport. A shape reconstruction method for electromagnetic tomography using adjoint fields and level sets. *Inverse Problems*, 16(5):1119–1156, October 2000. invited paper.

[26] Alvin W. Drake. *Fundamentals of Applied Probability Theory*. McGraw Hill, 1967.

[27] C. Henry Edwards and David E. Penney. *Elementary Differential Equations*. Prentice Hall, forth edition, 2000.

[28] H. Engl, M. Hanke, and A. Neubauer. *Regularization of Inverse Problems*. Kluwer Academic Publishers, 1996.

[29] H. W. Engl and W. Greever. Using the L-curve for determining optimal regularization parameters. *Numer. Math.*, 69:25–31, 1994.

[30] Leopold B. Felsen and Nathan Marcuvitz. *Radiation and Scattering of Waves*. The IEEE/OUP Series on Elecrtromagnetic Wave Theory. IEEE Press, 1994.

[31] H. Feng, D. A. Castañon, and W. C. Karl. Underground imaging based on edge-preserving regularization. In *Proc. IEEE International Conference on Information, Intelligence and Systems*, pages 460–464, Bethesda, MD, November 1999.

[32] H. Feng, D. A. Castañon, and W. C. Karl. Object-based reconstruction using coupled tomographic flows. In *Proc. IEEE International Conference on Image Processing*, Vacouver, BC, Canada, September 2000.

[33] Richard J. Gaudette, Dana H. Brooks, Charles A. DiMarzio, Misha E. Kilmer, Eric L. Miller, Tom Gaudette, and David Boas. A comparison study of linear reconstruction techniques for diffuse optical tomographic imaging of absorption coefficient. *Physics in Medicine and Biology*, 45(4):1051–1070, April 2000.

[34] Gene H. Golub, Michael Heath, and Grace Whaba. Generalzied cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2), 1975.

[35] Rafael C. Gonzalez and Richard E. Woods. *Digital Image Processing*. Prentice Hall, second edition, 2002.

[36] T. M. Habashy, W. C. Chew, and E. Y. Chow. Simultaneous reconstruction of permittivity and conductivity profiles in a radially inhomogeneous slab. *Radio Sci.*, 21(4):635–645, July–August 1986.

[37] T. M. Habashy and R. Mittra. On some inverse methods in electromagnetics. *Journal of Electromagnetic Waves and Applications*, 1(1):25–58, 1987.

[38] Tarek M. Habashy, Edward Y. Chow, and Donald G. Dudley. Profile inversion using the renormalized source-type integral equation approach. *IEEE Trans. on Antennas and Propagation*, 38(5):668–682, May 1990.

[39] M. Hanke. Limitations of the L-curve method in ill-posed problems. *BIT*, 36:287–301, 1996.

[40] P. C. Hansen and D. P. O'Leary. The use of the L-curve in the regularization of discrete ill-posed problems. *SIAM J. Scientific and Statistical Computing*, 14:1487–1503, 1993.

[41] Per Christian Hansen. Analysis of discrete ill-posed problems by means of the L-curve. *SIAM Review*, 34(4):561–580, December 1992.

[42] Per Christian Hansen. *Rank-Deficient and Discrete Illl-Posed Problems: Numerical Aspects of Linear Inversion*. SIAM, 1998.

[43] R. F. Harrington. *Field Computations by Moment Methods*. Macmillan, New York, 1968.

[44] Frank Hettlich. Frechet derivatives in inverse obstacle scattering. *Inverse Problems*, 11:371–382, 1995.

[45] Jr. J. E. Dennis and Robert B. Schnabel. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Prentice Hall, Inc., Englewood Cliffs, New Jersey, 1983. Series in Computational Mathematics.

[46] D. C. Munson Jr., J. D. O'Brien, and W. K. Jenkins. A tomographic formulation of spotlight-mode synthetic aperture radar. *Proceedings of the IEEE*, 1983.

[47] Avinash C. Kak and Malcolm Slaney. *Principles of Computerized Tomographic Imaging*. IEEE Press, Piscataway, N.J., 1987.

[48] Misha E. Kilmer. Cauchy-like preconditioners for 2-Dimensional ill-posed problems. *SIAM Journal of Matrix Analysis and Applications*, 20(3):777–799, 1999.

[49] Andreas Kirsch. *An Introduction to the Mathematical Theory of Inverse Problems*, volume 120 of *Applied Mathematical Sciences*. Springer-Verlag, 1996.

[50] Jin Au Kong. *Electromagnetic Fields*. John Wiley and Sons, 1986.

[51] Rainer Kress. Numerical methods in inverse acoustic obstacle cattering. In David Colton, Richard Ewing, and William Rundell, editors, *Inverse Problems in Partial Differential Equations*, chapter Chapter 5, pages 61–72. SIAM, 1990.

[52] Ranier Kress. *Linear Integral Equations*. Springer-Verlag, New York, 1989.

[53] K. T. Ladas and A. J. Devaney. Iterative methods in geophysical diffraction tomography. *Inverse Problems*, 8:119, 1992.

[54] A. Litman, D. Lesselier, and F. Santosa. Reconstruction of a two-dimensional binary obstacle by controlled evolution of a level-set. *Inverse Problems*, 14:685–706, 1998.

[55] David G. Luenberger. *Optimization by Vector Space Methods*. John Wiley and Sons, 1997.

[56] A. Mandelis. Theory of photothermal-wave diffraction and interference in condensed media. *J. Opt. Soc. Am.*, A6(298), 1989.

[57] Andreas Mandelis. *Diffusion Wave-Fields*. Springer-Verlag, New York, 2001.

[58] Dimitris G. Manolakis, Vinay K. Ingle, and Stephen M. Kogon. *Statistical and Adaptive Signal Processing*. McGraw Hill, 2000.

[59] Peyman Milanfar, William C. Karl, and Alan S. Willsky. Reconstructing binary polygonal objects from projections: A statistical view. *CVGIP: Graphical Models and Image Processing*, 56(5):371–391, September 1994.

[60] Peyman Milanfar, George C. Verghese, W. Clem Karl, and Alan S. Willsky. Reconstructing polygons from moments with connections to array processing. *IEEE Trans. Signal Processing*, 43(2):432–443, February 1995.

[61] Eric L. Miller. Statistically based methods for anomaly characterization in images from observations of scattered radiation. *IEEE Trans. on Image Processing*, 8(1):92–101, January 1999.

[62] Eric L. Miller, Misha E. Kilmer, and Carey M. Rappaport. A new shape-based method for object localization and characterization from scattered field data. *IEEE Trans. on Geoscience and Remote Sensing*, 38(4):1682–1696, July 2000. invited paper.

[63] Eric L. Miller, Lena Nicolaides, and Andreas Mandelis. Nonlinear inverse scattering methods for thermal wave slice tomography: A wavelet domain approach. *Journal of the Optical Society of America (A)*, 15(6):1545–1556, June 1998.

[64] Eric L. Miller, Ibrahim Yavuz, Lena Nicolaides, and Andreas Mandelis. An adaptive, multiscale inverse scattering approach to photothermal depth profilometry. *Circuits, Systems, and Signal Processing*, 19(4):339–363, August 2000. special issue on Advanced Signal/Image Restoration.

[65] M. Moghaddam, W.C. Chew, and M. Oristaglio. Comparison of the Born iterative method and Tarantola's method for an electromagnetic time-domain inverse problem. *International Journal of Imagin Systems and Technology*, 3:318–333, 1991.

[66] John E. Molyneux and Alan Witten. Impedance tomography: imaging algorithms for geophysical applications. *Inverse Problems*, 10:655–667, 1994.

[67] V. A. Morozov. On the solution of functional equations by the method of regularization. *Soviet Math. Dokl.*, 7:414–417, 1966.

[68] D.C. Munson and R.L. Visentin. A signal processing view of strip-mapping synthetic aperature radar. *IEEE Trans. on ASSP*, 36:2131–2147, December 1988.

[69] F. Natterer and Frank Wubbeling. *Mathematical Methods in Image Reconstruction*. SIAM, 2001.

[70] Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer Series in Operations Research. Springer-Verlag, 1999.

[71] Clifford Nolan and Margaret Cheney. Synthetic aperture inversion. *Inverse Problems*, to appear. Available at http://www.rpi.edu/ cheney/downloads.html.

[72] Stephen J. Norton. Iterative inverse scattering algorithms: Methods of computing Frechet derivatives. *Journal of the Acoustic Society of America*, 106:2653, 1999.

[73] A. A. Oberai and M. Malhotraand P. M. Pinsky. On thee implementation of the Dirichlet-to-Neumann map for iterative solution of the Helmholtz equation. *Journal of Applied Numerical Methods*, 27(4):443–464, 1998.

[74] M. A. O'Leary, D. A. Boas, B. Chance, and A. G. Yodh. Experimental images of heterogeneous turbid media by frequency-domain diffusing-photon tomography. *Optics Letters*, 20(5), March 1 1995.

[75] Athanasios Papoulis. *Probability, Random Variables, and Stochastic Processes*. McGraw Hill, second edition, 1984.

[76] William H. Press, Brian P. Flannery, Saul A. Teukolsky, and William T. Vetterling. *Numerical Recipes in C: The art of scientific computing*. Cambridge University Press, second edition, 1993.

[77] John G. Proakis and Dimitris G. Manolakis. *Digital Signal Processing: Principles, Algorithms and Applications*. Prentice Hall, third edition, 1995.

[78] David J. Rossi and Alan S. Willsky. Reconstruction from projections based on detection and estimation of objects–parts I and II: Performance analysis and robustness analysis. *IEEE Trans. on ASSP*, ASSP-32(4):886–906, August 1984.

[79] Yousef Saad. *Iterative Methods for Sparse Linear Systems*. PWS, Boston, 1996.

[80] Matthew N. O. Sadiku. *Numerical Techniques in Electromagnetics*. CRC Press, 2000.

[81] Fadil Santosa. A level-set approach for inverse problems involving obstacles. In *ESAIM: Control, Optimization, and Calculus of Variations*, pages 17–33. 1996.

[82] Samuli Siltanen, Jennifer Mueller, and David Isaacson. An implementation of the reconstruction algorithm of A Nachman for the 2D inverse conductivity problems. *Inverse Problems*, 16:681–699, 2000.

[83] Mehrdad Soumekh. *Synthetic Aperture Radar Signal Processing with MATLAB Algoirthms*. Wiley-Interscience, 1999.

[84] Henry Stark and Yongyi Yang. *Vector Space Projection: A Numerical Approach to Signal and Image Processing, Neural networks and Optics*. John Wiley and Sons, 1998.

[85] A. Taflove. *Computational Electrodynamics:The Finite Difference Time Domain Method*. Artech House, Boston, 1995.

[86] C. R. Vogel. Non-convergence of the L-curve regularization parameter selection method. *Inverse Problems*, 12:535–547, 1996.

[87] C. R. Vogel and M. E. Oman. Fast, robust total variation-based reconstruction of noisy, blurred images. *IEEE Trans. Image Process.*, 7(7):813–824, July 1998.

[88] Curtis R. Vogel. *Computational Methods for Inverse Problems*, volume FR23 of *Frontier in Applied Mathematics*. SIAM, 2002.

[89] J. C. Ye, K. J. Webb, C. A. Bouman, and R. P. Millane. Modified distorted Born iterative methods qith an approximate Frechet derivative for optical diffusion tomography. *Journal of the Optical Society of America-A*, 16(7):1814–1822, 1999.

[90] J. Zhang, R. L. Mackie, and T. R. Madden. 3-D resistivity forward modeling and inversion using conjugate grdients. *Geophysics*, 60:1313–1325, 1995.