Charles L. Byrne
Department of Mathematical Sciences
University of Massachusetts Lowell

Applied and Computational Linear Algebra: A First Course

To Eileen,
my wife for the last forty-three years.

My thanks to David Einstein, who read most of
an earlier version of this book
and made many helpful suggestions.

Contents

Preface

Those of us old enough to have first studied linear algebra in the 1960's remember a course devoted largely to proofs, devoid of applications and computation, full of seemingly endless discussion of the representation of linear transformations with respect to various bases, and concerned with matters that would not arise again in our mathematical education. With the growth of computer power and the discovery of powerful algorithms came the *digitization* of many problems previously analyzed solely in terms of functions of continuous variables. As it happened, I began my study of linear algebra in the fall of 1965, just as the two most important new algorithms in computational linear algebra appeared in print; the Cooley-Tukey Fast Fourier Transform (FFT) [101], and the Golub-Kahan method for computing the singular-value decomposition [149] would revolutionize applied linear algebra, but I learned of these more than a decade later. My experience was not at all unique; most of the standard linear algebra texts of the period, such as Cullen [105] and Hoffman and Kunze [168], ignored these advances.

Linear algebra, as we shall see, is largely the study of matrices, at least for the finite-dimensional cases. What connects the theory of matrices to applications are algorithms. Often the particular nature of the applications will prompt us to seek algorithms with particular properties; we then turn to the matrix theory to understand the workings of the algorithms. This book is intended as a text for a graduate course that focuses on applications of linear algebra and on the algorithms used to solve the problems that arise in those applications.

When functions of several continuous variables were approximated by finite-dimensional vectors, partial differential operators on these functions could be approximated by matrix multiplication. Images were represented in terms of grids of pixel values, that is, they became matrices, and then were vectorized into columns of numbers. Image processing then became the manipulation of these column vectors by matrix operations. This digitization meant that very large systems of linear equations now had to be dealt with. The need for fast algorithms to solve these large systems of linear equations turned linear algebra into a branch of applied and computational mathematics. Long forgotten topics in linear algebra, such as singular-value decomposition, were resurrected. Newly discovered algorithms, such as the

simplex method and the fast Fourier transform (FFT), revolutionized the field. As algorithms were increasingly applied to real-world data in real-world situations, the stability of these algorithms in the presence of noise became important. New algorithms emerged to answer the special needs of particular applications, and methods developed in other areas, such as likelihood maximization for statistical parameter estimation, found new application in reconstruction of medical and synthetic-aperture-radar (SAR) images.

The traditional topics of linear algebra, the geometry of Euclidean spaces, solving systems of linear equations and finding eigenvectors and eigenvalues, have not lost their importance, but now have a greater variety of roles to play. Orthogonal projections onto hyperplanes and convex sets form the building blocks for algorithms to design protocols for intensity-modulated radiation therapy. The unitary matrices that arise in discrete Fourier transformation are inverted quickly using the FFT, making essentially real-time magnetic-resonance imaging possible. In high-resolution radar and sonar, eigenvalues of certain matrices can tell us how many objects of interest are out there, while their eigenvectors can tell us where they are. Maximum-likelihood estimation of mixing probabilities lead to systems of linear equations to be solved to provide sub-pixel resolution of SAR images.

Part I

Preliminaries

Chapter 1

Introduction

1.1 Chapter Summary

This chapter introduces some of the topics to be considered in this course.

1.2 Overview of this Course

We shall focus here on applications that require the solution of systems of linear equations, often subject to constraints on the variables. These systems are typically large and sparse, that is, the entries of the matrices are predominantly zero. Transmission and emission tomography provide good examples of such applications. Fourier-based methods, such as filtered back-projection and the Fast Fourier Transform (FFT), are the standard tools for these applications, but statistical methods involving likelihood maximization are also employed. Because of the size of these problems and the nature of the constraints, iterative algorithms are essential.

Because the measured data is typically insufficient to specify a single unique solution, optimization methods, such as least-squares, likelihood maximization, and entropy maximization, are often part of the solution process. In the companion text "A First Course in Optimization", we present the fundamentals of optimization theory, and discuss *problems of optimization*, in which optimizing a function of one or several variables is the primary goal. Here, in contrast, our focus is on *problems of inference*,

optimization is not our primary concern, and optimization is introduced to overcome the non-uniqueness of possible solutions.

1.3 Solving Systems of Linear Equations

Many of the problems we shall consider involve solving, as least approximately, systems of linear equations. When an exact solution is sought and the number of equations and the number of unknowns are small, methods such as Gauss elimination can be used. It is common, in applications such as medical imaging, to encounter problems involving hundreds or even thousands of equations and unknowns. It is also common to prefer inexact solutions to exact ones, when the equations involve noisy, measured data. Even when the number of equations and unknowns is large, there may not be enough data to specify a unique solution, and we need to incorporate prior knowledge about the desired answer. Such is the case with medical tomographic imaging, in which the images are artificially discretized approximations of parts of the interior of the body.

1.4 Imposing Constraints

The iterative algorithms we shall investigate begin with an initial guess x^0 of the solution, and then generate a sequence $\{x^k\}$, converging, in the best cases, to our solution. When we use iterative methods to solve optimization problems, subject to constraints, it is necessary that the limit of the sequence $\{x^k\}$ of iterates obey the constraints, but not that each of the x^k do. An iterative algorithm is said to be an *interior-point method* if each vector x^k obeys the constraints. For example, suppose we wish to minimize $f(x)$ over all x in $\mathbb{R}^J$ having non-negative entries; an interior-point iterative method would have x^k non-negative for each k.

1.5 Operators

Most of the iterative algorithms we shall study involve an *operator*, that is, a function $T : \mathbb{R}^J \to \mathbb{R}^J$. The algorithms begin with an initial guess,

x^0, and then proceed from x^k to $x^{k+1} = Tx^k$. Ideally, the sequence $\{x^k\}$ converges to the solution to our optimization problem. To minimize the function $f(x)$ using a gradient descent method with fixed step-length α, for example, the operator is

$$Tx = x - \alpha \nabla f(x).$$

In problems with non-negativity constraints our solution x is required to have non-negative entries x_j. In such problems, the *clipping* operator T, with $(Tx)_j = \max\{x_j, 0\}$, plays an important role.

A subset C of $\mathbb{R}^J$ is *convex* if, for any two points in C, the line segment connecting them is also within C. As we shall see, for any x outside C, there is a point c within C that is closest to x; this point c is called the *orthogonal projection* of x onto C, and we write $c = P_C x$. Operators of the type $T = P_C$ play important roles in iterative algorithms. The clipping operator defined previously is of this type, for C the non-negative orthant of $\mathbb{R}^J$, that is, the set

$$\mathbb{R}^J_+ = \{x \in \mathbb{R}^J | x_j \geq 0,\ j = 1, ..., J\}.$$

1.6 Acceleration

For problems involving many variables, it is important to use algorithms that provide an acceptable approximation of the solution in a reasonable amount of time. For medical tomography image reconstruction in a clinical setting, the algorithm must reconstruct a useful image from scanning data in the time it takes for the next patient to be scanned, which is roughly fifteen minutes. Some of the algorithms we shall encounter work fine on small problems, but require far too much time when the problem is large. Figuring out ways to speed up convergence is an important part of iterative optimization. One approach we shall investigate in some detail is the use of *block-iterative* or *partial gradient* methods.

Chapter 2

An Overview of Applications

2.1 Chapter Summary

The theory of linear algebra, applications of that theory, and the associated computations are the three threads that weave their way through this course. In this chapter we present an overview of the applications we shall study in more detail later.

2.2 Transmission Tomography

Although transmission tomography (TT) is commonly associated with medical diagnosis, it has scientific uses, such as determining the sound-speed profile in the ocean, industrial uses, such as searching for faults in girders, mapping the interior of active volcanos, and security uses, such as the scanning of cargo containers for nuclear material. Previously, when people spoke of a "CAT scan" they usually meant x-ray transmission tomography, although the term is now used by lay people to describe any of the several scanning modalities in medicine, including single-photon emission computed tomography (SPECT), positron emission tomography (PET), ultrasound, and magnetic resonance imaging (MRI).

2.2.1 Brief Description

Computer-assisted tomography (CAT) scans have revolutionized medical practice. One example of CAT is transmission tomography. The goal here is to image the spatial distribution of various matter within the body, by estimating the distribution of radiation attenuation. At least in theory, the data are line integrals of the function of interest.

In transmission tomography, radiation, usually x-ray, is transmitted through the object being scanned. The object of interest need not be a living human being; King Tut has received a CAT-scan and industrial uses of transmission scanning are common. Recent work [235] has shown the practicality of using cosmic rays to scan cargo for hidden nuclear material; tomographic reconstruction of the scattering ability of the contents can reveal the presence of shielding. Because of their ability to penetrate granite, cosmic rays are being used to obtain transmission-tomographic three-dimensional images of the interior of active volcanos, to measure the size of the magma column and help predict the size and occurrence of eruptions.

In the simplest formulation of transmission tomography, the beams are

assumed to travel along straight lines through the object, the initial intensity of the beams is known and the intensity of the beams, as they exit the object, is measured for each line. The goal is to estimate and image the x-ray attenuation function, which correlates closely with the spatial distribution of attenuating material within the object. Unexpected absence of attenuation can indicate a broken bone, for example.

As the x-ray beam travels along its line through the body, it is weakened by the attenuating material it encounters. The reduced intensity of the exiting beam provides a measure of how much attenuation the x-ray encountered as it traveled along the line, but gives no indication of where along that line it encountered the attenuation; in theory, what we have learned is the integral of the attenuation function along the line. It is only by repeating the process with other beams along other lines that we can begin to localize the attenuation and reconstruct an image of this non-negative attenuation function. In some approaches, the lines are all in the same plane and a reconstruction of a single slice through the object is the goal; in other cases, a fully three-dimensional scanning occurs. The word "tomography" itself comes from the Greek "tomos", meaning part or slice; the word "atom" was coined to describe something supposed to be "without parts".

2.2.2 The Theoretical Problem

In theory, we will have the integral of the attenuation function along every line through the object. The *Radon Transform* is the operator that assigns to each attenuation function its integrals over every line. The mathematical problem is then to invert the Radon Transform, that is, to recapture the attenuation function from its line integrals. Is it always possible to determine the attenuation function from its line integrals? Yes. One way to show this is to use the Fourier transform to prove what is called the *Central Slice Theorem*. The reconstruction is then inversion of the Fourier transform; various methods for such inversion rely on frequency-domain filtering and back-projection.

2.2.3 The Practical Problem

Practise, of course, is never quite the same as theory. The problem, as we have described it, is an over-simplification in several respects, the main one being that we never have all the line integrals. Ultimately, we will construct a discrete image, made up of finitely many pixels. Consequently, it is reasonable to assume, from the start, that the attenuation function to be estimated is well approximated by a function that is constant across small squares (or cubes), called pixels (or voxels), and that the goal is to determine these finitely many pixel values.

2.2.4 The Discretized Problem

When the problem is discretized in this way, different mathematics begins to play a role. The line integrals are replaced by finite sums, and the problem can be viewed as one of solving a large number of linear equations, subject to side constraints, such as the non-negativity of the pixel values. The Fourier transform and the Central Slice Theorem are still relevant, but in discrete form, with the fast Fourier transform (FFT) playing a major role in discrete filtered back-projection methods. This approach provides fast reconstruction, but is limited in other ways. Alternatively, we can turn to iterative algorithms for solving large systems of linear equations, subject to constraints. This approach allows for greater inclusion of the physics into the reconstruction, but can be slow; accelerating these iterative reconstruction algorithms is a major concern, as is controlling sensitivity to noise in the data.

2.2.5 Mathematical Tools

As we just saw, Fourier transformation in one and two dimensions, and frequency-domain filtering are important tools that we need to discuss in some detail. In the discretized formulation of the problem, periodic convolution of finite vectors and its implementation using the fast Fourier transform play major roles. Because actual data is always finite, we consider the issue of under-determined problems that allow for more than one answer, and the need to include prior information to obtain reasonable reconstructions. Under-determined problems are often solved using optimization, such as maximizing the entropy or minimizing the norm of the image, subject to the data as constraints. Constraints are often described mathematically using the notion of convex sets. Finding an image satisfying several sets of constraints can often be viewed as finding a vector in the intersection of convex sets, the so-called *convex feasibility problem* (CFP).

2.3 Emission Tomography

Unlike transmission tomography, emission tomography (ET) is used only with living beings, principally humans and small animals. Although this modality was initially used to uncover pathologies, it is now used to study normal functioning, as well. In emission tomography, including positron emission tomography (PET) and single photon emission tomography (SPECT), the patient inhales, swallows, or is injected with, chemicals to which radioactive material has been chemically attached [263]. The

chemicals are designed to accumulate in that specific region of the body we wish to image. For example, we may be looking for tumors in the abdomen, weakness in the heart wall, or evidence of brain activity in a selected region. In some cases, the chemicals are designed to accumulate more in healthy regions, and less so, or not at all, in unhealthy ones. The opposite may also be the case; tumors may exhibit greater avidity for certain chemicals. The patient is placed on a table surrounded by detectors that count the number of emitted photons. On the basis of where the various counts were obtained, we wish to determine the concentration of radioactivity at various locations throughout the region of interest within the patient.

Although PET and SPECT share some applications, their uses are generally determined by the nature of the chemicals that have been designed for this purpose, as well as by the half-life of the radionuclides employed. Those radioactive isotopes used in PET generally have half-lives on the order of minutes and must be manufactured on site, adding to the expense of PET. The isotopes used in SPECT have half-lives on the order of many hours, or even days, so can be manufactured off-site and can also be used in scanning procedures that extend over some appreciable period of time.

2.3.1 Coincidence-Detection PET

In a typical PET scan to detect tumors, the patient receives an injection of glucose, to which a radioactive isotope of fluorine, ^{18}F, has been chemically attached. The radionuclide emits individual positrons, which travel, on average, between 4 mm and 2.5 cm (depending on their kinetic energy) before encountering an electron. The resulting annihilation releases two gamma-ray photons that then proceed in essentially opposite directions. Detection in the PET case means the recording of two photons at nearly the same time at two different detectors. The locations of these two detectors then provide the end points of the line segment passing, more or less, through the site of the original positron emission. Therefore, each possible pair of detectors determines a *line of response* (LOR). When a LOR is recorded, it is assumed that a positron was emitted somewhere along that line. The PET data consists of a chronological list of LOR that are recorded. Because the two photons detected at either end of the LOR are not detected at exactly the same time, the time difference can be used in *time-of-flight* PET to further localize the site of the emission to a smaller segment of perhaps 8 cm in length.

2.3.2 Single-Photon Emission Tomography

Single-photon computed emission tomography (SPECT) is similar to PET and has the same objective: to image the distribution of a radionuclide, such as technetium ^{99m}Tc, within the body of the patient. In SPECT

the radionuclide employed emits single gamma-ray photons, which then travel through the body of the patient and, in some fraction of the cases, are detected. Detections in SPECT correspond to individual sensor locations outside the body. The data in SPECT are the photon counts at each of the finitely many detector locations. Unlike PET, in SPECT lead collimators are placed in front of the gamma-camera detectors to eliminate photons arriving at oblique angles. While this helps us narrow down the possible sources of detected photons, it also reduces the number of detected photons and thereby decreases the signal-to-noise ratio.

2.3.3 The Line-Integral Model for PET and SPECT

To solve the reconstruction problem we need a model that relates the count data to the radionuclide density function. A somewhat unsophisticated, but computationally attractive, model is taken from transmission tomography: to view the count at a particular detector as the line integral of the radionuclide density function along the line from the detector that is perpendicular to the camera face. The count data then provide many such line integrals and the reconstruction problem becomes the familiar one of estimating a function from noisy measurements of line integrals. Viewing the data as line integrals allows us to use the Fourier transform in reconstruction. The resulting *filtered back-projection* (FBP) algorithm is a commonly used method for medical imaging in clinical settings.

The line-integral model for PET assumes a fixed set of possible LOR, with most LOR recording many emissions. Another approach is *list-mode* PET, in which detections are recording as they occur by listing the two end points of the associated LOR. The number of potential LOR is much higher in list-mode, with most of the possible LOR being recording only once, or not at all [173, 216, 61].

2.3.4 Problems with the Line-Integral Model

It is not really accurate, however, to view the photon counts at the detectors as line integrals. Consequently, applying filtered back-projection to the counts at each detector can lead to distorted reconstructions. There are at least three degradations that need to be corrected before FBP can be successfully applied [181]: attenuation, scatter, and spatially dependent resolution.

In the SPECT case, as in most such inverse problems, there is a trade-off to be made between careful modeling of the physical situation and computational tractability. The FBP method slights the physics in favor of computational simplicity and speed. In recent years, iterative methods, such as the *algebraic reconstruction technique* (ART), its multiplicative variant, MART, the expectation maximization maximum likelihood (MLEM

or EMML) method, and the rescaled block-iterative EMML (RBI-EMML), that incorporate more of the physics have become competitive.

2.3.5 The Stochastic Model: Discrete Poisson Emitters

In iterative reconstruction we begin by *discretizing* the problem; that is, we imagine the region of interest within the patient to consist of finitely many tiny squares, called *pixels* for two-dimensional processing or cubes, called *voxels* for three-dimensional processing. We imagine that each pixel has its own level of concentration of radioactivity and these concentration levels are what we want to determine. Proportional to these concentration levels are the average rates of emission of photons. To achieve our goal we must construct a model that relates the measured counts to these concentration levels at the pixels. The standard way to do this is to adopt the model of *independent Poisson emitters*. Any Poisson-distributed random variable has a mean equal to its variance. The *signal-to-noise ratio* (SNR) is usually taken to be the ratio of the mean to the standard deviation, which, in the Poisson case, is then the square root of the mean. Consequently, the Poisson SNR increases as the mean value increases, which points to the desirability (at least, statistically speaking) of higher dosages to the patient.

2.3.6 Reconstruction as Parameter Estimation

The goal is to reconstruct the distribution of radionuclide intensity by estimating the pixel concentration levels. The pixel concentration levels can be viewed as parameters and the data are instances of random variables, so the problem looks like a fairly standard parameter estimation problem of the sort studied in beginning statistics. One of the basic tools for statistical parameter estimation is likelihood maximization, which is playing an increasingly important role in medical imaging. There are several problems, however.

One problem is that the number of parameters is quite large, as large as the number of data values, in most cases. Standard statistical parameter estimation usually deals with the estimation of a handful of parameters. Another problem is that we do not quite know the relationship between the pixel concentration levels and the count data. The reason for this is that the probability that a photon emitted from a given pixel will be detected at a given detector will vary from one patient to the next, since whether or not a photon makes it from a given pixel to a given detector depends on the geometric relationship between detector and pixel, as well as what is in the patient's body between these two locations. If there are ribs or skull getting in the way, the probability of making it goes down. If there are just lungs, the probability goes up. These probabilities can change during the

scanning process, when the patient moves. Some motion is unavoidable, such as breathing and the beating of the heart. Determining good values of the probabilities in the absence of motion, and correcting for the effects of motion, are important parts of SPECT image reconstruction.

2.3.7 X-Ray Fluorescence Computed Tomography

X-ray fluorescence computed tomography (XFCT) is a form of emission tomography that seeks to reconstruct the spatial distribution of elements of interest within the body [191]. Unlike SPECT and PET, these elements need not be radioactive. Beams of synchrotron radiation are used to stimulate the emission of fluorescence x-rays from the atoms of the elements of interest. These fluorescence x-rays can then be detected and the distribution of the elements estimated and imaged. As with SPECT, attenuation is a problem; making things worse is the lack of information about the distribution of attenuators at the various fluorescence energies.

2.4 Magnetic Resonance Imaging

Protons have *spin*, which, for our purposes here, can be viewed as a charge distribution in the nucleus revolving around an axis. Associated with the resulting current is a *magnetic dipole moment* collinear with the axis of the spin. In elements with an odd number of protons, such as hydrogen, the nucleus itself will have a net magnetic moment. The objective in *magnetic resonance imaging* (MRI) is to determine the density of such elements in a volume of interest within the body. The basic idea is to use strong magnetic fields to force the individual spinning nuclei to emit signals that, while too weak to be detected alone, are detectable in the aggregate. The signals are generated by the precession that results when the axes of the magnetic dipole moments are first aligned and then perturbed.

In much of MRI, it is the distribution of hydrogen in water molecules that is the object of interest, although the imaging of phosphorus to study energy transfer in biological processing is also important. There is ongoing work using tracers containing fluorine, to target specific areas of the body and avoid background resonance. Because the magnetic properties of blood change when the blood is oxygenated, increased activity in parts of the brain can be imaged through *functional* MRI (fMRI). Non-radioactive isotopes of gadolinium are often injected as contrast agents because of their ability to modify certain parameters called the T1 relaxation times.

2.4.1 Alignment

In the absence of an external magnetic field, the axes of these magnetic dipole moments have random orientation, dictated mainly by thermal effects. When an external magnetic field is introduced, it induces a small fraction, about one in 10^5, of the dipole moments to begin to align their axes with that of the external magnetic field. Only because the number of protons per unit of volume is so large do we get a significant number of moments aligned in this way. A strong external magnetic field, about $20,000$ times that of the earth's, is required to produce enough alignment to generate a detectable signal.

2.4.2 Precession

When the axes of the aligned magnetic dipole moments are perturbed, they begin to precess, like a spinning top, around the axis of the external magnetic field, at the *Larmor frequency*, which is proportional to the intensity of the external magnetic field. If the magnetic field intensity varies spatially, then so does the Larmor frequency. Each precessing magnetic dipole moment generates a signal; taken together, they contain information about the density of the element at the various locations within the body. As we shall see, when the external magnetic field is appropriately chosen, a Fourier relationship can be established between the information extracted from the received signal and this density function.

2.4.3 Slice Isolation

When the external magnetic field is the *static field*, then the Larmor frequency is the same everywhere. If, instead, we impose an external magnetic field that varies spatially, then the Larmor frequency is also spatially varying. This external field is now said to include a *gradient field*.

2.4.4 Tipping

When a magnetic dipole moment is given a component out of its axis of alignment, it begins to precess around its axis of alignment, with frequency equal to its Larmor frequency. To create this off-axis component, we apply a *radio-frequency field* (rf field) for a short time. The effect of imposing this rf field is to tip the aligned magnetic dipole moment axes away from the axis of alignment, initiating precession. The dipoles that have been tipped ninety degrees out of their axis of alignment generate the strongest signal.

2.4.5 Imaging

The information we seek about the proton density function is contained within the received signal. By carefully adding gradient fields to the external field, we can make the Larmor frequency spatially varying, so that each frequency component of the received signal contains a piece of the information we seek. The proton density function is then obtained through Fourier transformations. Fourier-transform estimation and extrapolation techniques play a major role in this rapidly expanding field [157].

2.4.6 The Line-Integral Approach

By appropriately selecting the gradient field and the radio-frequency field, it is possible to create a situation in which the received signal comes primarily from dipoles along a given line in a preselected plane. Performing an FFT of the received signal gives us line integrals of the density function along lines in that plane. In this way, we obtain the three-dimensional Radon transform of the desired density function. The Central Slice Theorem for this case tells us that, in theory, we have the Fourier transform of the density function.

2.4.7 Phase Encoding

In the line-integral approach, the line-integral data is used to obtain values of the Fourier transform of the density function along lines through the origin in Fourier space. It would be more convenient for the FFT if we have Fourier-transform values on the points of a rectangular grid. We can obtain this by selecting the gradient fields to achieve *phase encoding*.

2.4.8 A New Application

A recent article [262] in The Boston Globe describes a new application of MRI, as a guide for the administration of ultra-sound to kill tumors and perform bloodless surgery. In MRI-guided focused ultra-sound, the sound waves are focused to heat up the regions to be destroyed and real-time MRI imaging shows the doctor where this region is located and if the sound waves are having the desired effect. The use of this technique in other areas is also being studied: to open up the blood-brain barrier to permit chemo-therapy for brain cancers; to cure hand tremors, chronic pain, and some effects of stroke, epilepsy, and Parkinson's disease; and to remove uterine fibroids.

2.5 Intensity Modulated Radiation Therapy

A fairly recent addition to the list of applications using linear algebra and the geometry of Euclidean space is *intensity modulated radiation therapy* (IMRT). Although it is not actually an imaging problem, intensity modulated radiation therapy is an emerging field that involves some of the same mathematical techniques used to solve the medical imaging problems discussed previously, particularly methods for solving the convex feasibility problem.

2.5.1 Brief Description

In IMRT beamlets of radiation with different intensities are transmitted into the body of the patient. Each voxel within the patient will then absorb a certain dose of radiation from each beamlet. The goal of IMRT is to direct a sufficient dosage to those regions requiring the radiation, those that are designated *planned target volumes* (PTV), while limiting the dosage received by the other regions, the so-called *organs at risk* (OAR).

2.5.2 The Problem and the Constraints

The intensities and dosages are obviously non-negative quantities. In addition, there are *implementation constraints*; the available treatment machine will impose its own requirements, such as a limit on the difference in intensities between adjacent beamlets. In dosage space, there will be a lower bound on the acceptable dosage delivered to those regions designated as the PTV, and an upper bound on the acceptable dosage delivered to those regions designated as the OAR. The problem is to determine the intensities of the various beamlets to achieve these somewhat conflicting goals.

2.5.3 Convex Feasibility and IMRT

The CQ algorithm [62, 63] is an iterative algorithm for solving the split feasibility problem. Because it is particularly simple to implement in many cases, it has become the focus of recent work in IMRT. In [84] Censor *et al.* extend the CQ algorithm to solve what they call the *multiple-set split feasibility problem* (MSSFP) . In the sequel [82] it is shown that the constraints in IMRT can be modeled as inclusion in convex sets and the extended CQ algorithm is used to determine dose intensities for IMRT that satisfy both dose constraints and radiation-source constraints.

One drawback to the use of x-rays in radiation therapy is that they continue through the body after they have encountered their target. A re-

cent technology, proton-beam therapy, directs a beam of protons at the target. Since the protons are heavy, and have mass and charge, their trajectories can be controlled in ways that x-ray trajectories cannot be. The new proton center at Massachusetts General Hospital in Boston is one of the first to have this latest technology. As with most new and expensive medical procedures, there is some debate going on about just how much of an improvement it provides, relative to other methods.

2.6 Array Processing

Passive SONAR is used to estimate the number and direction of distant sources of acoustic energy that have generated sound waves propagating through the ocean. An array, or arrangement, of sensors, called *hydrophones*, is deployed to measure the incoming waveforms over time and space. The data collected at the sensors is then processed to provide estimates of the waveform parameters being sought. In active SONAR, the party deploying the array is also the source of the acoustic energy, and what is sensed are the returning waveforms that have been reflected off of distant objects. Active SONAR can be used to map the ocean floor, for example. Radar is another active array-processing procedure, using reflected radio waves instead of sound to detect distant objects. Radio astronomy uses array processing and the radio waves emitted by distant sources to map the heavens.

To illustrate how array processing operates, consider Figure 2.1. Imagine a source of acoustic energy sufficiently distant from the line of sensors that the incoming wavefront is essentially planar. As the peaks and troughs of the wavefronts pass over the array of sensors, the measurements at the sensors give the elapsed time between a peak at one sensor and a peak at the next sensor, thereby giving an indication of the angle of arrival.

In practice, of course, there are multiple sources of acoustic energy, so each sensor receives a superposition of all the plane-wave fronts from all directions. because the sensors are spread out in space, what each receives is slightly different from what its neighboring sensors receive, and this slight difference can be exploited to separate the spatially distinct components of the signals. What we seek is the function that describes how much energy came from each direction.

When we describe the situation mathematically, using the wave equation, we find that what is received at each sensor is a value of the Fourier transform of the function we want. Because we have only finitely many sensors, we have only finitely many values of this Fourier transform. So, we

have the problem of estimating a function from finitely many values of its Fourier transform.

2.7 A Word about Prior Information

An important point to keep in mind when applying linear-algebraic methods to measured data is that, while the data is usually limited, the information we seek may not be lost. Although processing the data in a reasonable way may suggest otherwise, other processing methods may reveal that the desired information is still available in the data. Figure 2.2 illustrates this point.

The original image on the upper right of Figure 2.2 is a discrete rectangular array of intensity values simulating a slice of a head. The data was obtained by taking the two-dimensional discrete Fourier transform of the original image, and then discarding, that is, setting to zero, all these spatial frequency values, except for those in a smaller rectangular region around the origin. The problem then is under-determined. A minimum two-norm solution would seem to be a reasonable reconstruction method.

The minimum two-norm solution is shown on the lower right. It is calculated simply by performing an inverse discrete Fourier transform on the array of modified discrete Fourier transform values. The original image has relatively large values where the skull is located, but the minimum two-norm reconstruction does not want such high values; the norm involves the sum of squares of intensities, and high values contribute disproportionately to the norm. Consequently, the minimum two-norm reconstruction chooses instead to conform to the measured data by spreading what should be the skull intensities throughout the interior of the skull. The minimum two-norm reconstruction does tell us something about the original; it tells us about the existence of the skull itself, which, of course, is indeed a prominent feature of the original. However, in all likelihood, we would already know about the skull; it would be the interior that we want to know about.

Using our knowledge of the presence of a skull, which we might have obtained from the minimum two-norm reconstruction itself, we construct the prior estimate shown in the upper left. Now we use the same data as before, and calculate a minimum weighted two-norm solution, using as the weight vector the reciprocals of the values of the prior image. This minimum weighted two-norm reconstruction is shown on the lower left; it is clearly almost the same as the original image. The calculation of the minimum weighted two-norm solution can be done iteratively using the ART algorithm, as discussed in [238].

When we weight the skull area with the inverse of the prior image, we allow the reconstruction to place higher values there without having much of an effect on the overall weighted norm. In addition, the reciprocal weighting in the interior makes spreading intensity into that region costly, so the interior remains relatively clear, allowing us to see what is really present there.

When we try to reconstruct an image from limited data, it is easy to assume that the information we seek has been lost, particularly when a reasonable reconstruction method fails to reveal what we want to know. As this example, and many others, show, the information we seek is often still in the data, but needs to be brought out in a more subtle way.

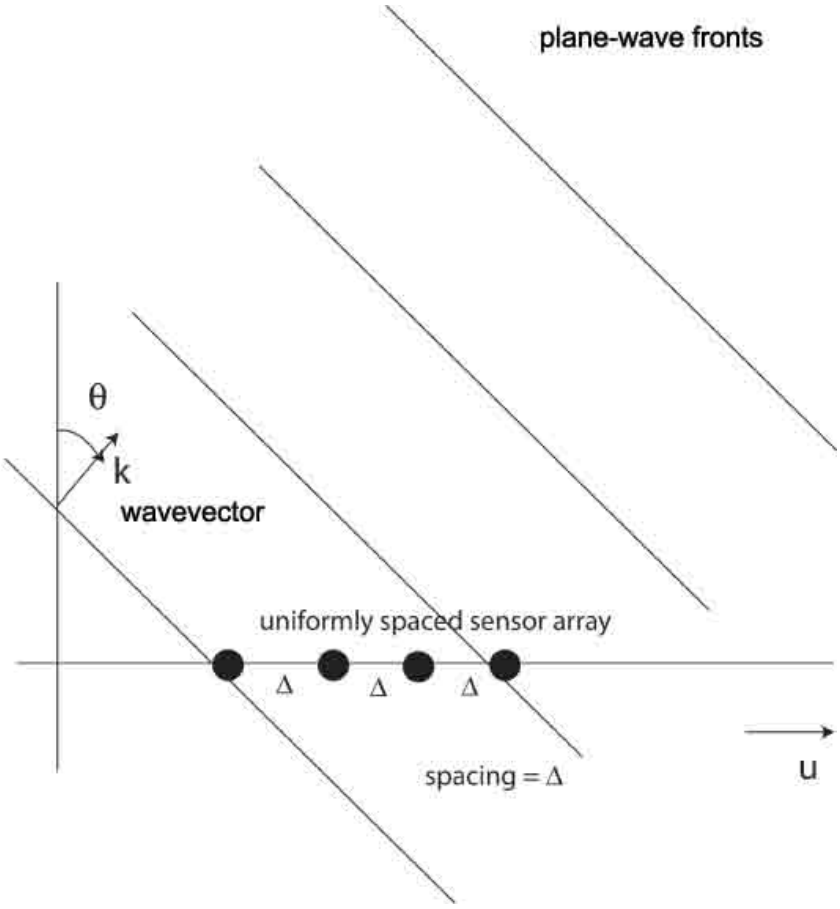

FIGURE 2.1: A uniform line array sensing a plane-wave field.

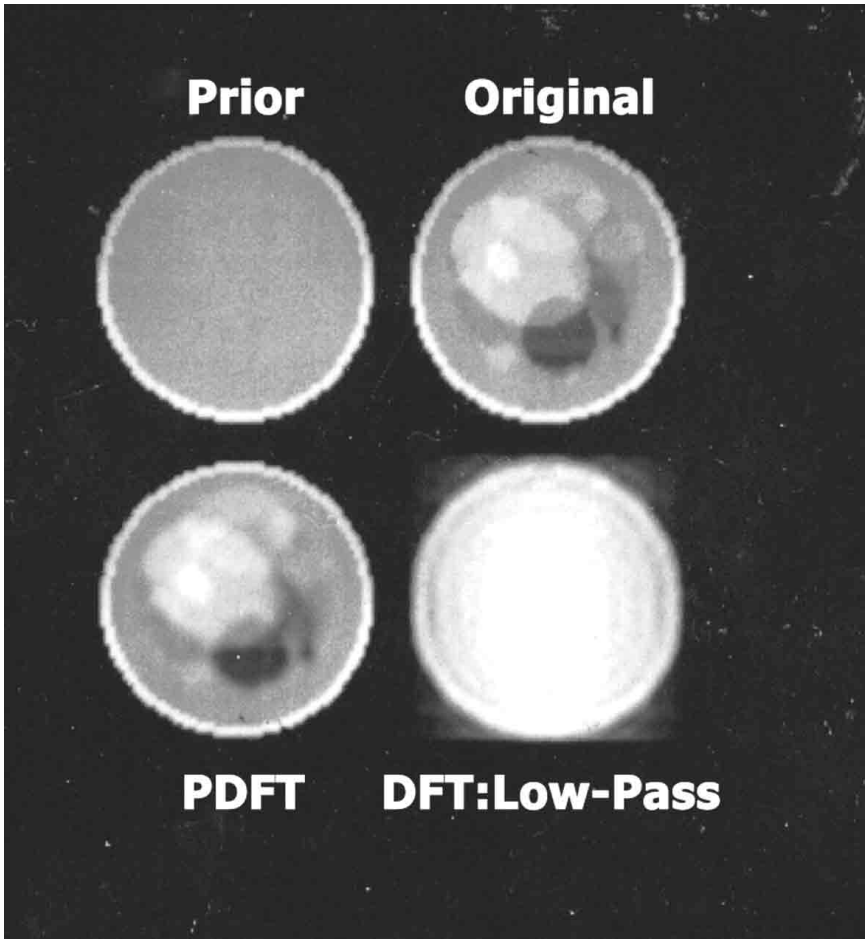

FIGURE 2.2: Extracting information in image reconstruction.

Chapter 3

A Little Matrix Theory

3.1 Chapter Summary

In this chapter we review the fundamentals of matrix algebra.

3.2 Vector Spaces

Linear algebra is the study of *vector spaces* and *linear transformations*. It is not simply the study of matrices, although matrix theory takes up most of linear algebra.

It is common in mathematics to consider abstraction, which is simply a means of talking about more than one thing at the same time. A vector space V is an abstract algebraic structure defined using axioms. There are many examples of vector spaces, such as the sets of real or complex numbers themselves, the set of all polynomials, the set of row or column vectors of a given dimension, the set of all infinite sequences of real or complex numbers, the set of all matrices of a given size, and so on. The beauty of an abstract approach is that we can talk about all of these, and much more, all at once, without being specific about which example we mean.

A vector space is a set whose members are called *vectors*, on which there are two algebraic operations, called *scalar multiplication* and *vector addition*. As in any axiomatic approach, these notions are intentionally abstract. A vector is defined to be a member of a vector space, nothing more. Scalars are a bit more concrete, in that scalars are almost always real or complex numbers, although sometimes, but not in this book, they are members of an unspecified finite field. The operations themselves are not explicitly defined, except to say that they behave according to certain axioms, such as associativity and distributivity.

If v is a member of a vector space V and α is a scalar, then we denote by αv the scalar multiplication of v by α. If w is also a member of V, then we denote by $v+w$ the vector addition of v and w. The following properties serve to define a vector space, with u, v, and w denoting arbitrary members of V and α and β arbitrary scalars:

- 1. $v + w = w + v$;
- 2. $u + (v + w) = (u + v) + w$;
- 3. there is a unique "zero vector" , denoted 0, such that, for every v, $v + 0 = v$;
- 4. for each v there is a unique vector $-v$ such that $v + (-v) = 0$;
- 5. $1v = v$, for all v;
- 6. $(\alpha\beta)v = \alpha(\beta v)$;
- 7. $\alpha(v + w) = \alpha v + \alpha w$;
- 8. $(\alpha + \beta)v = \alpha v + \beta v$.

Ex. 3.1 *Show that, if $z + z = z$, then z is the zero vector.*

Ex. 3.2 *Prove that $0v = 0$, for all $v \in V$, and use this to prove that $(-1)v = -v$ for all $v \in V$. Hint: use Exercise 3.1.*

We then write

$$w - v = w + (-v) = w + (-1)v,$$

for all v and w.

If $u^1, ..., u^N$ are members of V and $c_1, ..., c_N$ are scalars, then the vector

$$x = c_1 u^1 + c_2 u^2 + ... + c_N u^N$$

is called a *linear combination* of the vectors $u^1, ..., u^N$, with coefficients $c_1, ..., c_N$.

If W is a subset of a vector space V, then W is called a *subspace* of V if W is also a vector space for the same operations. What this means is simply that when we perform scalar multiplication on a vector in W, or when we add vectors in W, we always get members of W back again. Another way to say this is that W is *closed to linear combinations.*

When we speak of subspaces of V we do not mean to exclude the case of $W = V$. Note that V is itself a subspace, but not a *proper subspace* of V. Every subspace must contain the zero vector, 0; the smallest subspace of V is the subspace containing only the zero vector, $W = \{0\}$.

Ex. 3.3 *Show that, in the vector space $V = \mathbb{R}^2$, the subset of all vectors whose entries sum to zero is a subspace, but the subset of all vectors whose entries sum to one is not a subspace.*

Ex. 3.4 *Let V be a vector space, and W and Y subspaces of V. Show that the union of W and Y, written $W \cup Y$, is also a subspace if and only if either $W \subseteq Y$ or $Y \subseteq W$.*

We often refer to things like $\begin{bmatrix} 1 & 2 & 0 \end{bmatrix}$ as vectors, although they are but one example of a certain type of vector. For clarity, in this book we shall call such an object a *real row vector of dimension three* or a *real row three-vector.* Similarly, we shall call $\begin{bmatrix} 3i \\ -1 \\ 2+i \\ 6 \end{bmatrix}$ a *complex column vector of dimension four* or a *complex column four-vector.* For notational convenience, whenever we refer to something like a real three-vector or a complex four-vector, we shall always mean that they are columns, rather than rows. The space of real (column) N-vectors will be denoted $\mathbb{R}^N$, while the space of complex (column) N vectors is $\mathbb{C}^N$.

Shortly after beginning a discussion of vector spaces, we arrive at the

notion of the size or dimension of the vector space. A vector space can be finite dimensional or infinite dimensional. The spaces $\mathbb{R}^N$ and $\mathbb{C}^N$ have dimension N; not a big surprise. The vector spaces of all infinite sequences of real or complex numbers are infinite dimensional, as is the vector space of all real or complex polynomials. If we choose to go down the path of finite dimensionality, we very quickly find ourselves talking about matrices. If we go down the path of infinite dimensionality, we quickly begin to discuss convergence of infinite sequences and sums, and find that we need to introduce norms, which takes us into functional analysis and the study of Hilbert and Banach spaces. In this course we shall consider only the finite dimensional vector spaces, which means that we shall be talking mainly about matrices.

3.3 Matrix Algebra

A system $Ax = b$ of linear equations is called a *complex system*, or a *real system* if the entries of A, x and b are complex, or real, respectively. Note that when we say that the entries of a matrix or a vector are complex, we do not intend to rule out the possibility that they are real, but just to open up the possibility that they are not real.

3.3.1 Matrix Operations

If A and B are real or complex M by N and N by K matrices, respectively, then the product $C = AB$ is defined as the M by K matrix whose entry C_{mk} is given by

$$C_{mk} = \sum_{n=1}^{N} A_{mn} B_{nk}. \tag{3.1}$$

If x is an N-dimensional column vector, that is, x is an N by 1 matrix, then the product $b = Ax$ is the M-dimensional column vector with entries

$$b_m = \sum_{n=1}^{N} A_{mn} x_n. \tag{3.2}$$

Ex. 3.5 *Show that, for each $k = 1, ..., K$, $\mathrm{Col}_k(C)$, the kth column of the matrix $C = AB$, is*

$$\mathrm{Col}_k(C) = A\mathrm{Col}_k(B).$$

It follows from this exercise that, for given matrices A and C, every column of C is a linear combination of the columns of A if and only if there is a third matrix B such that $C = AB$.

For any N, we denote by I the N by N identity matrix with entries $I_{n,n} = 1$ and $I_{m,n} = 0$, for $m, n = 1, ..., N$ and $m \neq n$. The size of I is always to be inferred from the context.

The matrix $A^\dagger$ is the *conjugate transpose* of the matrix A, that is, the N by M matrix whose entries are

$$(A^\dagger)_{nm} = \overline{A_{mn}} \tag{3.3}$$

When the entries of A are real, $A^\dagger$ is just the *transpose* of A, written A^T.

Definition 3.1 *A square matrix S is* symmetric *if $S^T = S$ and* Hermitian *if $S^\dagger = S$.*

Definition 3.2 *A square matrix S is* normal *if $S^\dagger S = SS^\dagger$.*

Ex. 3.6 *Let $C = AB$. Show that $C^\dagger = B^\dagger A^\dagger$.*

Ex. 3.7 *Let D be a diagonal matrix such that $D_{mm} \neq D_{nn}$ if $m \neq n$. Show that if $BD = DB$ then B is a diagonal matrix.*

Ex. 3.8 *Prove that, if $AB = BA$ for every N by N matrix A, then $B = cI$, for some constant c.*

3.3.2 Matrix Inverses

We begin with the definition of invertibility.

Definition 3.3 *A square matrix A is said to be* invertible*, or to be a* non-singular *matrix if there is a matrix B such that*

$$AB = BA = I$$

where I is the identity matrix of the appropriate size. There can be at most one such matrix B for a given A. Then $B = A^{-1}$, the inverse *of A.*

Note that, in this definition, the matrices A and B must commute.

Proposition 3.1 *The inverse of a square matrix A is unique; that is, if $AB = BA = I$ and $AC = CA = I$, then $B = C = A^{-1}$.*

Ex. 3.9 *Prove Proposition 3.1.*

The following proposition shows that invertibility follows from an apparently weaker condition.

Proposition 3.2 *If A is square and there exist matrices B and C such that $AB = I$ and $CA = I$, then $B = C = A^{-1}$ and A is invertible.*

Ex. 3.10 *Prove Proposition 3.2.*

Later in this chapter, after we have discussed the concept of rank of a matrix, we will improve Proposition 3.2; a square matrix A is invertible if and only if there is a matrix B with $AB = I$, and, for any (possibly non-square) A, if there are matrices B and C with $AB = I$ and $CA = I$ (where the two I may possibly be different in size), then A must be square and invertible.

The 2 by 2 matrix $S = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ has an inverse

$$S^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

whenever the *determinant* of S, $\det(S) = ad - bc$ is not zero. More generally, associated with every complex square matrix is the complex number called its determinant, which is obtained from the entries of the matrix using formulas that can be found in any text on linear algebra. The significance of the determinant is that the matrix is invertible if and only if its determinant is not zero. This is of more theoretical than practical importance, since no computer can tell when a number is precisely zero. A matrix A that is not square cannot have an inverse, but does have a *pseudo-inverse*, which can be found using the singular-value decomposition.

Note that, if A is invertible, then $Ax = 0$ can happen only when $x = 0$. We shall show later, using the notion of the rank of a matrix, that the converse is also true: a square matrix A with the property that $Ax = 0$ only when $x = 0$ must be invertible.

3.3.3 The Sherman-Morrison-Woodbury Identity

In a number of applications, stretching from linear programming to radar tracking, we are faced with the problem of computing the inverse of a slightly modified version of a matrix B, when the inverse of B itself has already been computed. For example, when we use the simplex algorithm in linear programming, the matrix B consists of some, but not all, of the columns of a larger matrix A. At each step of the simplex algorithm, a new B_{new} is formed from $B = B_{\text{old}}$ by removing one column of B and replacing it with another column taken from A.

Then B_{new} differs from B in only one column. Therefore

$$B_{\text{new}} = B_{\text{old}} - uv^T, \tag{3.4}$$

where u is the column vector that equals the old column minus the new one, and v is the column of the identity matrix corresponding to the column of B_{old} being altered. The inverse of B_{new} can be obtained fairly easily from the inverse of B_{old} using the Sherman-Morrison-Woodbury Identity:

The Sherman-Morrison-Woodbury Identity: If $v^T B^{-1} u \neq 1$, then

$$(B - uv^T)^{-1} = B^{-1} + \alpha^{-1}(B^{-1}u)(v^T B^{-1}), \tag{3.5}$$

where

$$\alpha = 1 - v^T B^{-1} u.$$

Ex. 3.11 *Let B be invertible and $v^T B^{-1} u = 1$. Show that $B - uv^T$ is not invertible. Show that Equation (3.5) holds, if $v^T B^{-1} u \neq 1$.*

3.4 Bases and Dimension

The related notions of a basis and of linear independence are fundamental in linear algebra.

3.4.1 Linear Independence and Bases

As we shall see shortly, the *dimension* of a *finite-dimensional* vector space will be defined as the number of members of any basis. Obviously, we first need to see what a basis is, and then to convince ourselves that if a vector space V has a basis with N members, then every basis for V has N members.

Definition 3.4 *The* span *of a collection of vectors $\{u^1, ..., u^N\}$ in V is the set of all vectors x that can be written as linear combinations of the u^n; that is, for which there are scalars $c_1, ..., c_N$, such that*

$$x = c_1 u^1 + ... + c_N u^N. \tag{3.6}$$

Definition 3.5 *A collection of vectors $\{w^1, ..., w^N\}$ in V is called a* spanning set *for a subspace W if the set W is their span.*

Definition 3.6 *A subspace W of a vector space V is called* finite dimensional *if it is the span of a finite set of vectors from V. The whole space V is then finite dimensional if it is the span of a finite set of vectors.*

The assertion in the following proposition may seem obvious, but the proof, which the reader is asked to supply as Exercise 3.12, is surprisingly subtle.

Proposition 3.3 *Let V be a finite dimensional vector space and W a subspace of V. Then W is also finite dimensional.*

Ex. 3.12 *Prove Proposition 3.3.*

This definition tells us what it means to be finite dimensional, but does not tell us what *dimension* means, nor what the actual dimension of a finite dimensional subset is; for that we need the notions of *linear independence* and *basis*.

Definition 3.7 *A collection of vectors $\mathcal{U} = \{u^1, ..., u^N\}$ in V is* linearly independent *if there is no choice of scalars $\alpha_1, ..., \alpha_N$, not all zero, such that*

$$0 = \alpha_1 u^1 + ... + \alpha_N u^N. \tag{3.7}$$

Ex. 3.13 *Show that the following are equivalent:*

- *1. the set $\mathcal{U} = \{u^1, ..., u^N\}$ is linearly independent;*
- *2. $u^1 \neq 0$ and no u^n is a linear combination of the members of $\mathcal{U}$ that precede it in the list;*
- *3. no u^n is a linear combination of the other members of $\mathcal{U}$.*

Definition 3.8 *A collection of vectors $\mathcal{U} = \{u^1, ..., u^N\}$ in V is called a* basis *for a subspace W if the collection is linearly independent and W is their span.*

Ex. 3.14 *Show that*

- *1. if $\mathcal{U} = \{u^1, ..., u^N\}$ is a spanning set for W, then $\mathcal{U}$ is a basis for W if and only if, after the removal of any one member, $\mathcal{U}$ is no longer a spanning set for W; and*
- *2. if $\mathcal{U} = \{u^1, ..., u^N\}$ is a linearly independent set in W, then $\mathcal{U}$ is a basis for W if and only if, after including in $\mathcal{U}$ any new member from W, $\mathcal{U}$ is no longer linearly independent.*

Ex. 3.15 *Prove that every finite dimensional vector space that is not just the zero vector has a basis.*

3.4.2 Dimension

We turn now to the task of showing that every basis for a finite dimensional vector space has the same number of members. That number will then be used to define the dimension of that space.

Suppose that W is a subspace of V, that $\mathcal{W} = \{w^1, ..., w^N\}$ is a spanning set for W, and $\mathcal{U} = \{u^1, ..., u^M\}$ is a linearly independent subset of W. Beginning with w^1, we augment the set $\{u^1, ..., u^M\}$ with w^j if w^j is not in the span of the u^m and the w^k previously included. At the end of this process, we have a linearly independent spanning set, and therefore, a basis, for W (Why?). Similarly, beginning with w^1, we remove w^j from the set $\{w^1, ..., w^N\}$ if w^j is a linear combination of the w^k, $k = 1, ..., j-1$. In this way we obtain a linearly independent set that spans W, hence another basis for W. The following lemma will allow us to prove that all bases for a subspace W have the same number of elements.

Lemma 3.1 *Let $\mathcal{W} = \{w^1, ..., w^N\}$ be a spanning set for a subspace W of V, and $\mathcal{U} = \{u^1, ..., u^M\}$ a linearly independent subset of W. Then $M \leq N$.*

Proof: Suppose that $M > N$. Let $B_0 = \mathcal{W} = \{w^1, ..., w^N\}$. To obtain the set B_1, form the set $C_1 = \{u^1, w^1, ..., w^N\}$ and remove the first member of C_1 that is a linear combination of members of C_1 that occur to its left in the listing; since u^1 has no members to its left, it is not removed. Since $\mathcal{W}$ is a spanning set, $u^1 \neq 0$ is a linear combination of the members of $\mathcal{W}$, so that some member of $\mathcal{W}$ is a linear combination of u^1 and the members of $\mathcal{W}$ to the left of it in the list; remove the first member of $\mathcal{W}$ for which this is true.

We note that the set B_1 is a spanning set for W and has N members. Having obtained the spanning set B_k, with N members and whose first k members are $u^k, ..., u^1$, we form the set $C_{k+1} = B_k \cup \{u^{k+1}\}$, listing the members so that the first $k+1$ of them are $\{u^{k+1}, u^k, ..., u^1\}$. To get the set B_{k+1} we remove the first member of C_{k+1} that is a linear combination of the members to its left; there must be one, since B_k is a spanning set, and so u^{k+1} is a linear combination of the members of B_k. Since the set $\mathcal{U}$ is linearly independent, the member removed is from the set $\mathcal{W}$. Continuing in this fashion, we obtain a sequence of spanning sets $B_1, ..., B_N$, each with N members. The set B_N is $B_N = \{u^N, ..., u^1\}$ and u^{N+1} must then be a linear combination of the members of B_N, which contradicts the linear independence of $\mathcal{U}$. ∎

Corollary 3.1 *Every basis for a subspace W has the same number of elements.*

Definition 3.9 *The* dimension of a subspace W, *denoted* $\dim(W)$, *is the number of elements in any basis.*

Ex. 3.16 *Let V be a finite dimensional vector space and W any subspace of V. Show that* $\dim(W)$ *cannot exceed* $\dim(V)$.

3.4.3 Rank of a Matrix

We rely on the following lemma to define the rank of a matrix.

Lemma 3.2 *For any matrix A, the maximum number of linearly independent rows equals the maximum number of linearly independent columns.*

Proof: Suppose that A is an M by N matrix, and that $K \leq N$ is the maximum number of linearly independent columns of A. Select K linearly independent columns of A and use them as the K columns of an M by K matrix U. Since every column of A must be a linear combination of these K selected ones, there is a K by N matrix B such that $A = UB$; see the discussion that follows Exercise 3.5. From $A^\dagger = B^\dagger U^\dagger$ we conclude that every column of $A^\dagger$ is a linear combination of the K columns of the matrix $B^\dagger$. Therefore, there can be at most K linearly independent columns of $A^\dagger$. ∎

Definition 3.10 *The* rank *of A, written rank(A), is the maximum number of linearly independent rows or of linearly independent columns of A.*

Ex. 3.17 *Let u and v be two non-zero N-dimensional complex column vectors. Show that the rank of the N by N matrix $uv^\dagger$ is one.*

Ex. 3.18 *Show that the rank of a matrix $C = AB$ is never greater than the smaller of the rank of A and the rank of B. Can it ever be strictly less than the smaller of these two numbers?*

Ex. 3.19 *Show that rank(A+B) is never greater than the sum of rank(A) and rank(B).*

Definition 3.11 *An M by N matrix A is said to have* full rank *or to be a* full-rank matrix *if the rank of A is the minimum of M and N.*

Proposition 3.4 *A square matrix is invertible if and only if it has full rank.*

Ex. 3.20 *Prove Proposition 3.4.*

Corollary 3.2 *A square matrix A is invertible if and only if there is a matrix B such that $AB = I$.*

Corollary 3.3 *A square matrix A is invertible if and only if there is a matrix G such that AG is invertible.*

Corollary 3.4 *If A and B are square matrices and $C = AB$ is invertible, then both A and B are invertible.*

Definition 3.12 *An M by N matrix A is said to have* left inverse *B if B is an N by M matrix such that $BA = I_N$, the N by N identity matrix. Similarly, A is said to have a* right inverse *C if C is an N by M matrix such that $AC = I_M$, the M by M identity matrix.*

Ex. 3.21 *Let A be an M by N matrix. When does A have a left inverse? When does it have a right inverse?*

Ex. 3.22 *Let A and B be M by N matrices, P an invertible M by M matrix, and Q an invertible N by N matrix, such that $B = PAQ$, that is, the matrices A and B are* equivalent*. Show that the rank of B is the same as the rank of A. Hint: show that A and AQ have the same rank.*

3.5 Representing a Linear Transformation

Let V and W be vector spaces. A function $T : V \to W$ is called a *linear transformation* if

$$T(\alpha u + \beta v) = \alpha T(u) + \beta T(v),$$

for all scalars α and β and all u and v in V. For notational convenience we often write simply Tu instead of $T(u)$. When both V and W are finite-dimensional a linear transformation can be represented by a matrix, which is why we say that there is a close relationship between abstract linear algebra and matrix theory.

Let $\mathcal{A} = \{a^1, a^2, ..., a^N\}$ be a basis for the finite-dimensional complex vector space V. Now that the basis for V is specified, there is a natural association, an *isomorphism*, between V and the vector space $\mathbb{C}^N$ of N-dimensional column vectors with complex entries. Any vector v in V can be written as

$$v = \sum_{n=1}^{N} \gamma_n a^n. \tag{3.8}$$

The column vector $\gamma = (\gamma_1, ..., \gamma_N)^T$ is uniquely determined by v and the basis $\mathcal{A}$ and we denote it by $[v]_{\mathcal{A}}$. Notice that the ordering of the list of

members of $\mathcal{A}$ matters, so we shall always assume that the ordering has been fixed.

Let W be a second finite-dimensional vector space, and let T be any linear transformation from V to W. Let $\mathcal{B} = \{b^1, b^2, ..., b^M\}$ be a basis for W. For $n = 1, ..., N$, let

$$Ta^n = A_{1n}b^1 + A_{2n}b^2 + ... + A_{Mn}b^M. \tag{3.9}$$

Then the M by N matrix A having the A_{mn} as entries is said to *represent* T, with respect to the bases $\mathcal{A}$ and $\mathcal{B}$, and we write $A = [T]^{\mathcal{B}}_{\mathcal{A}}$.

Ex. 3.23 *Show that* $[Tv]_{\mathcal{B}} = A[v]_{\mathcal{A}}$.

Ex. 3.24 *Suppose that V, W and Z are vector spaces, with bases $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$, respectively. Suppose also that T is a linear transformation from V to W and U is a linear transformation from W to Z. Let A represent T with respect to the bases $\mathcal{A}$ and $\mathcal{B}$, and let B represent U with respect to the bases $\mathcal{B}$ and $\mathcal{C}$. Show that the matrix BA represents the linear transformation UT with respect to the bases $\mathcal{A}$ and $\mathcal{C}$.*

3.6 The Geometry of Euclidean Space

We denote by $\mathbb{R}^N$ the real Euclidean space consisting of all N-dimensional column vectors $x = (x_1, ..., x_N)^T$ with real entries x_j; here the superscript T denotes the transpose of the 1 by N matrix (or, row vector) $(x_1, ..., x_N)$. We denote by $\mathbb{C}^N$ the space of all N-dimensional column vectors with complex entries. For x in $\mathbb{C}^N$ we denote by $x^\dagger$ the N-dimensional row vector whose entries are the complex conjugates of the entries of x.

3.6.1 Dot Products

For $x = (x_1, ..., x_N)^T$ and $y = (y_1, ..., y_N)^T$ in $\mathbb{C}^N$, the dot product $x \cdot y$ is defined to be

$$x \cdot y = \sum_{n=1}^{N} x_n \overline{y_n}. \tag{3.10}$$

Note that we can write

$$x \cdot y = y^\dagger x, \tag{3.11}$$

where juxtaposition indicates matrix multiplication. The 2-norm, or *Euclidean norm*, or *Euclidean length*, of x is

$$||x||_2 = \sqrt{x \cdot x} = \sqrt{x^\dagger x}. \tag{3.12}$$

The *Euclidean distance* between two vectors x and y in $\mathbb{C}^N$ is $||x - y||_2$. These notions also apply to vectors in $\mathbb{R}^N$.

In subsequent chapters we shall consider norms for vectors other than the two-norm. However, for the remainder of this chapter all vector norms are the two-norm.

The spaces $\mathbb{R}^N$ and $\mathbb{C}^N$, along with their dot products, are examples of a finite-dimensional Hilbert space.

Definition 3.13 *Let V be a real or complex vector space. The scalar-valued function $\langle u, v\rangle$ is called an* inner product *on V if the following four properties hold, for all u, w, and v in V, and all scalars c:*

$$\langle u + w, v\rangle = \langle u, v\rangle + \langle w, v\rangle; \tag{3.13}$$

$$\langle cu, v\rangle = c\langle u, v\rangle; \tag{3.14}$$

$$\langle v, u\rangle = \overline{\langle u, v\rangle}; \tag{3.15}$$

and

$$\langle u, u\rangle \geq 0, \tag{3.16}$$

with equality in Inequality (3.16) if and only if $u = 0$.

Once we have an inner product on the vector space V we also have a norm, denoted $\|\cdot\|_2$ defined by

$$\|u\|_2^2 = \langle u, u\rangle.$$

The dot products on $\mathbb{R}^N$ and $\mathbb{C}^N$ are examples of inner products. The properties of an inner product are precisely the ones needed to prove Cauchy's Inequality, which then holds for any inner product. We shall favor the dot product notation $u \cdot v$ for the inner product of vectors in $\mathbb{R}^N$ or $\mathbb{C}^N$, although we shall occasionally use the matrix multiplication form, $v^\dagger u$ or the inner product notation $\langle u, v\rangle$.

Ex. 3.25 *Show that, for any real number λ, we have*

$$\|\lambda x + (1-\lambda)y\|^2 + \lambda(1-\lambda)\|x - y\|^2 = \lambda\|x\|^2 + (1-\lambda)\|y\|^2. \tag{3.17}$$

We may conclude from Exercise 3.25 that, for any α in the interval $(0,1)$ and x not equal to y, we have

$$\|\alpha x + (1-\alpha)y\|^2 < \alpha\|x\|^2 + (1-\alpha)\|y\|^2, \tag{3.18}$$

so that the square of the norm is a strictly convex function.

Definition 3.14 *A collection of vectors* $\{u^1, ..., u^N\}$ *in an inner product space* V *is called* orthonormal *if* $||u^n||_2 = 1$*, for all* n*, and* $\langle u^m, u^n\rangle = 0$*, for* $m \neq n$*.*

3.6.2 Cauchy's Inequality

Cauchy's Inequality, also called the Cauchy-Schwarz Inequality, tells us that

$$|\langle x, y\rangle| \leq ||x||_2||y||_2, \tag{3.19}$$

with equality if and only if $y = \alpha x$, for some scalar α. The Cauchy-Schwarz Inequality holds for any inner product.

A simple application of Cauchy's inequality gives us

$$||x + y||_2 \leq ||x||_2 + ||y||_2; \tag{3.20}$$

this is called the *Triangle Inequality*. We say that the vectors x and y are *mutually orthogonal* if $\langle x, y\rangle = 0$.

The *Parallelogram Law* is an easy consequence of the definition of the 2-norm:

$$||x + y||_2^2 + ||x - y||_2^2 = 2||x||_2^2 + 2||y||_2^2. \tag{3.21}$$

It is important to remember that Cauchy's Inequality and the Parallelogram Law hold only for the 2-norm.

3.7 Vectorization of a Matrix

When the complex M by N matrix A is stored in the computer it is usually *vectorized*; that is, the matrix

$$A = \begin{bmatrix} A_{11} & A_{12} & \dots & A_{1N} \\ A_{21} & A_{22} & \dots & A_{2N} \\ \vdots & & & \\ A_{M1} & A_{M2} & \dots & A_{MN} \end{bmatrix}$$

becomes

$$\text{vec}(A) = (A_{11}, A_{21}, ..., A_{M1}, A_{12}, A_{22}, ..., A_{M2}, ..., A_{MN})^T.$$

Definition 3.15 *The* trace *of a square matrix A, abbreviated* $\text{tr}(A)$*, is the sum of the entries on its main diagonal.*

It can be shown that

$$\text{tr}(ABC) = \text{tr}(CAB) = \text{tr}(BCA), \tag{3.22}$$

but it is not generally true that $\text{tr}(ABC) = \text{tr}(BAC)$.

Ex. 3.26 • **a)** *Show that the complex dot product* $\text{vec}(A)\cdot \text{vec}(B) = \text{vec}(B)^\dagger \text{vec}(A)$ *can be obtained by*

$$\text{vec}(A)\cdot \text{vec}(B) = \text{trace}\,(AB^\dagger) = tr(AB^\dagger).$$

We can therefore use the trace to define an inner product between matrices: $< A, B >= \text{trace}\,(AB^\dagger)$.

• **b)** *Show that* $\text{trace}\,(AA^\dagger) \geq 0$ *for all A, so that we can use the trace to define a norm on matrices:* $||A||_F^2 = \text{trace}\,(AA^\dagger)$*. This norm is the Frobenius norm*

3.8 Solving Systems of Linear Equations

In this section we discuss systems of linear equations, Gaussian elimination, and the notions of basic and non-basic variables.

3.8.1 Row-Reduction

One approach to solving systems of linear equations is to use elementary row operations to convert the original system to another system with the same solutions.

Definition 3.16 *There are three types of* elementary row operations*. The first is to multiply a given row by a scalar. The second is to switch two rows. The third is to add to a given row some multiple of another row.*

Definition 3.17 *An M by N matrix B is said to be in* row-reduced echelon form *if the following conditions hold:*

- *1. the first non-zero entry of any row is a one;*
- *2. in any column containing one of these "first non-zero" ones, the remaining entries are zero;*
- *3. all zero rows come at the bottom; and*
- *4. if $j < k$ then the column containing the first non-zero entry of the jth row occurs before the column containing the first non-zero entry of the kth row.*

Lemma 3.3 *Any matrix A can be transformed into a matrix B in row-reduced echelon form using elementary row operations.*

Ex. 3.27 *Prove Lemma 3.3.*

Proposition 3.5 *Let A be an M by N matrix with rank R. Then there are invertible matrices P and Q such that PAQ is a diagonal matrix with the entries of the R by R identity matrix in the upper left corner and all the rest of the entries equal to zero.*

Proof: We know that any matrix A can be transformed to row-reduced echelon form using row operations, or, equivalently, by multiplying A on the left by elementary matrices. The proof follows by applying the same reasoning to $A^\dagger$. ∎

Proposition 3.6 *Let A be an arbitrary M by N matrix and B the matrix in row-reduced echelon form obtained from A. There is a non-zero solution of the system of linear equations $Ax = 0$ if and only if B has fewer than N non-zero rows.*

Ex. 3.28 *Prove Proposition 3.6.*

Corollary 3.5 *If A is M by N and $M < N$, then there is a non-zero x with $Ax = 0$.*

Ex. 3.29 *Prove Corollary 3.5.*

Ex. 3.30 *Let $\mathcal{W} = \{w^1, ..., w^N\}$ be a spanning set for a subspace W in $\mathbb{R}^K$, and $\mathcal{U} = \{u^1, ..., u^M\}$ a linearly independent subset of W. Let A be the K by M matrix whose columns are the vectors u^m and B the K by N matrix whose columns are the w^n. Then there is an N by M matrix D such that $A = BD$ (Why?). Prove Lemma 3.1 for this case by showing that, if $M > N$, then there is a non-zero vector x with $Dx = 0$.*

Definition 3.18 *Let A be an M by N matrix. The* null space *of A, denoted NS(A), is the set of all x such that $Ax = 0$. The* nullity *of A, denoted $n(A)$, is the dimension of its null space.*

Proposition 3.7 *Let A be an N by N matrix with rank $J < N$. Then there are $N - J$ linearly independent solutions of the system $Ax = 0$, and the null space of A has dimension $N - J$.*

Ex. 3.31 *Prove Proposition 3.7.*

3.8.2 Row Operations as Matrix Multiplications

Suppose that we want to apply a row operation to the M by N matrix A. We can first apply that row operation to the M by M identity matrix, to obtain the new matrix E, and then multiply A by E on the left. The matrix EA is exactly what we would have obtained if we had just performed the row operation on A directly. For example, to multiply the first row of A by k we could multiply A by the matrix $E_1(k)$, which is the identity matrix, except that the one in the first row is replaced by k.

If A is square and we are able to row reduce A to the identity matrix I, then there are matrices E_1, E_2,...,E_J such that

$$E_J E_{J-1} \cdots E_2 E_1 A = I.$$

It follows then that

$$E_J E_{J-1} \cdots E_2 E_1 = A^{-1}.$$

We can also use this approach to calculate the determinant of A.

3.8.3 Determinants

Associated with each square matrix A is a number, its determinant, denoted $\det(A)$. Most texts that discuss determinants define the concept by telling us how to compute it. There is a different way that is more interesting (see [105]).

We define the determinant to be a complex-valued function of square complex matrices having the following two properties:

- 1. $\det(AB) = \det(A)\det(B)$ for all compatible square matrices A and B;
- 2. the determinant of the matrix $E_1(k)$ is k, where $E_1(k)$ is as defined in the previous subsection.

Using only these two properties, we can prove the following concerning the effects of row operations on the determinant of A:

- 1. multiplying one row by k multiplies the determinant by k;
- 2. interchanging two rows changes the sign of the determinant;

- 3. adding to one row a multiple of another row has no effect on the determinant.

Ex. 3.32 *Prove these assertions concerning the effects of row operations on the determinant.*

Of course, it remains to be shown that such a function of square matrices exists. To show the existence of such a function it is sufficient to show how it may be calculated, for any given square matrix A. Once we have converted A to an upper triangular matrix using row operations we can calculate the determinant of A immediately, since the determinant of an upper triangular matrix can easily be shown to be the product of the entries along its main diagonal. If we prefer, we can use more row operations to convert A to row-reduced echelon form. If A is invertible, this reduction produces the identity matrix, with determinant equal to one. If A is not invertible, the row-reduced echelon form will have a zero row, so that the determinant is zero. After we have discussed Schur's Lemma we shall prove the following proposition.

Proposition 3.8 *A square matrix is invertible if and only if its determinant is not zero.*

3.8.4 Sylvester's Nullity Theorem

Recall that the nullity of a matrix A is $n(A)$, the dimension of its null space. The following is taken from [77].

Theorem 3.1 Sylvester's Nullity Theorem *Let A and B be M by N and N by J matrices, respectively. Then*

- *1. $n(AB) \leq n(A) + n(B)$;*
- *2. $n(AB) \geq n(A)$;*
- *3. $n(AB) \geq n(B)$, provided that $M \geq N$.*

Proof: Let R be $r(A)$, the rank of A. Select invertible matrices P and Q so that $PAQ = A^*$ has the entries of the R by R identity matrix in the upper left corner and zeros everywhere else. Set $B^* = Q^{-1}B$. Then A^*, B^*, and $A^*B^* = PAB$ are equivalent to, so have the same ranks and nullities as, A, B and AB, respectively.

The first R rows of A^*B^* are those of B^*, and the remaining $M - R$ ones are zero. The matrix B^* has $r(B^*) = r(B)$ linearly independent rows, of which at most $N - R$ do not appear in A^*B^*. Therefore, there must be at least $r(B) - (N - R) = r(A) + r(B) - N$ linearly independent rows in A^*B^*, and so $r(A^*B^*) \geq r(A) + r(B) - N$.

We know that $r(A) = N - n(A)$, $r(B) = J - n(B)$, and

$$r(AB) = r(A^*B^*) = J - n(A^*B^*) = J - n(AB).$$

Therefore,

$$J - n(AB) \geq N - n(A) + J - n(B) - N,$$

so that $n(AB) \leq n(A) + n(B)$.

The null space of A is a subspace of the null space of AB, so that $n(A) \leq n(AB)$.

Since $r(AB) \leq r(B)$, we have $n(B) \leq M - r(B) \leq n(AB)$, provided that $N \leq M$. ∎

3.8.5 Homogeneous Systems of Linear Equations

Consider the system of three linear equations in five unknowns given by

$$\begin{aligned} x_1 + 2x_2 \ &+ 2x_4 + x_5 = 0 \\ -x_1 - x_2 + x_3 + x_4 \ &= 0 \\ x_1 + 2x_2 - 3x_3 - x_4 - 2x_5 &= 0. \end{aligned} \tag{3.23}$$

This system can be written in matrix form as $Ax = 0$, with A the coefficient matrix

$$A = \begin{bmatrix} 1 & 2 & 0 & 2 & 1 \\ -1 & -1 & 1 & 1 & 0 \\ 1 & 2 & -3 & -1 & -2 \end{bmatrix}, \tag{3.24}$$

and $x = (x_1, x_2, x_3, x_4, x_5)^T$.

The standard approach to solving a system of M equations in N unknowns is to apply Gaussian elimination, to obtain a second, simpler, system with the same solutions. To avoid potential numerical difficulties, Gauss elimination may involve *row pivoting*, which means that when we are about to eliminate the variable x_k from the equations $k+1$ through M, we switch the kth row with the one below it that has the coefficient of x_k with the largest absolute value. In the example below we do not employ pivoting.

Using Gaussian elimination, we obtain the equivalent system of equations

$$\begin{aligned} x_1 - 2x_4 + x_5 &= 0 \\ x_2 + 2x_4 &= 0 \\ x_3 + x_4 + x_5 &= 0. \end{aligned} \tag{3.25}$$

The associated matrix is B, the row reduced echelon form matrix obtained from A:

$$B = \begin{bmatrix} 1 & 0 & 0 & -2 & 5 \\ 0 & 1 & 0 & 2 & 0 \\ 0 & 0 & 1 & 1 & 1 \end{bmatrix}. \tag{3.26}$$

From this simpler system we see that the variables x_4 and x_5 can be freely chosen, with the other three variables then determined by this system of equations. The variables x_4 and x_5 are then *independent*, the others *dependent*. The variables x_1, x_2 and x_3 are then called *basic variables*; note that this terminology is commonly used in linear programming, but has nothing to do with the notion of a basis. To obtain a basis of solutions we can let $x_4 = 1$ and $x_5 = 0$, obtaining the solution $x = (2, -2, -1, 1, 0)^T$, and then choose $x_4 = 0$ and $x_5 = 1$ to get the solution $x = (-1, 0, -1, 0, 1)^T$. Every solution to $Ax = 0$ is then a linear combination of these two solutions. Notice that which variables are basic and which are non-basic is somewhat arbitrary, in that we could have chosen as the non-basic variables any two whose columns are independent.

Having decided that x_4 and x_5 are the non-basic variables, we can write the original matrix A in block-matrix form as $A = \begin{bmatrix} B & C \end{bmatrix}$, where B is the square invertible matrix

$$B = \begin{bmatrix} 1 & 2 & 0 \\ -1 & -1 & 1 \\ 1 & 2 & -3 \end{bmatrix}, \tag{3.27}$$

and C is the matrix

$$C = \begin{bmatrix} 2 & 1 \\ 1 & 0 \\ -1 & -2 \end{bmatrix}. \tag{3.28}$$

With $x_B = (x_1, x_2, x_3)^T$ and $x_C = (x_4, x_5)^T$ the vector x can be written in concatenated form as a block matrix, that is,

$$x = \begin{bmatrix} x_B^T & x_C^T \end{bmatrix}^T = \begin{bmatrix} x_B \\ x_C \end{bmatrix}.$$

Now we can write

$$Ax = Bx_B + Cx_C = 0, \tag{3.29}$$

so that

$$x_B = -B^{-1}Cx_C. \tag{3.30}$$

3.8.6 Real and Complex Systems of Linear Equations

Any complex system can be converted to a real system in the following way. A complex matrix A can be written as $A = A_1 + iA_2$, where A_1 and A_2 are real matrices and $i = \sqrt{-1}$. Similarly, $x = x^1 + ix^2$ and $b = b^1 + ib^2$, where x^1, x^2, b^1 and b^2 are real vectors. Denote by $\tilde{A}$ the real matrix

$$\tilde{A} = \begin{bmatrix} A_1 & -A_2 \\ A_2 & A_1 \end{bmatrix}, \tag{3.31}$$

by $\tilde{x}$ the real vector

$$\tilde{x} = \begin{bmatrix} x^1 \\ x^2 \end{bmatrix}, \tag{3.32}$$

and by $\tilde{b}$ the real vector

$$\tilde{b} = \begin{bmatrix} b^1 \\ b^2 \end{bmatrix}. \tag{3.33}$$

Then x satisfies the system $Ax = b$ if and only if $\tilde{x}$ satisfies the system $\tilde{A}\tilde{x} = \tilde{b}$.

3.9 Under-Determined Systems of Linear Equations

Suppose that $Ax = b$ is a linear system of M equations in N unknowns, where $M < N$. Then we say that the system is *under-determined*. Typically, there will be an infinite number of solutions, although there need not be any solutions in this case (give an example). A standard procedure in such cases is to find that solution x having the smallest two-norm

$$||x||_2 = \sqrt{\sum_{n=1}^{N} |x_n|^2}.$$

As we shall see shortly, the *minimum two-norm* solution of $Ax = b$ is a vector of the form $x = A^\dagger z$, where $A^\dagger$ denotes the conjugate transpose of the matrix A. Then $Ax = b$ becomes $AA^\dagger z = b$. Typically, $(AA^\dagger)^{-1}$ will exist, and we get $z = (AA^\dagger)^{-1}b$, from which it follows that the minimum two-norm solution is $x = A^\dagger(AA^\dagger)^{-1}b$. When M and N are not too large, forming the matrix $AA^\dagger$ and solving for z is not prohibitively expensive or time-consuming. However, in image processing the vector x is often a vectorization of a two-dimensional (or even three-dimensional) image and

M and N can be on the order of tens of thousands or more. The ART algorithm gives us a fast method for finding the minimum two-norm solution without computing $AA^\dagger$.

We begin by describing the minimum two-norm solution of a consistent system $Ax = b$, starting with the fundamental *subspace decomposition* lemma.

Lemma 3.4 *For every x in $\mathbb{C}^N$ there are unique vectors $A^\dagger z$ in the range of $A^\dagger$ and w in the null space of A, such that $x = A^\dagger z + w$.*

Proof: The z that minimizes the function

$$f(z) = \frac{1}{2}\|x - A^\dagger z\|_2^2$$

satisfies the equation

$$0 = \nabla f(z) = A(x - A^\dagger z).$$

Then $w = x - A^\dagger z$ satisfies $Aw = 0$. Expanding $\|x\|^2 = \|A^\dagger z + w\|^2$ and using the fact that $Aw = 0$ we find that

$$\|x\|^2 = \|A^\dagger z\|^2 + \|w\|^2.$$

If we also had

$$x = A^\dagger \hat{z} + \hat{w},$$

with $A\hat{w} = 0$, then, writing

$$A^\dagger z = A^\dagger \hat{z} + \hat{w} - w,$$

we could conclude that

$$\|A^\dagger z\|^2 = \|A^\dagger \hat{z}\|^2 + \|\hat{w} - w\|^2.$$

But writing

$$A^\dagger \hat{z} = A^\dagger z + w - \hat{w},$$

we could also conclude that

$$\|A^\dagger \hat{z}\|^2 = \|A^\dagger z\|^2 + \|w - \hat{w}\|^2.$$

It follows then that $\hat{w} = w$ and that $A^\dagger \hat{z} = A^\dagger z$. ∎

Corollary 3.6 *For every M by N matrix A and every b in $\mathbb{C}^M$ there are unique vectors x in $\mathbb{C}^N$ and w in $\mathbb{C}^M$ such that $A^\dagger w = 0$ and $b = Ax + w$.*

Corollary 3.7 *An N by N matrix A is invertible if and only if $Ax = 0$ implies $x = 0$.*

Proof: If A is invertible and $Ax = 0$, then clearly we must have $x = 0$. Conversely, suppose that $Ax = 0$ only when $x = 0$. Then the null space of A is the subspace of $\mathbb{C}^N$ consisting only of the zero vector. Consequently, every vector in $\mathbb{C}^N$ lies in the column space of $A^\dagger$, so that N is the rank of $A^\dagger$, which is also the rank of A. So A has full rank and A must be invertible. ∎

Theorem 3.2 *The minimum two-norm solution of $Ax = b$ has the form $x = A^\dagger z$ for some M-dimensional complex vector z.*

Proof: If $Ax = b$ then $A(x + w) = b$ for all w in the null space of A. If $x = A^\dagger z$ and w is in the null space of A, then

$$\begin{aligned} ||x + w||_2^2 = ||A^\dagger z + w||_2^2 = (A^\dagger z + w)^\dagger (A^\dagger z + w) \\ = (A^\dagger z)^\dagger (A^\dagger z) + (A^\dagger z)^\dagger w + w^\dagger (A^\dagger z) + w^\dagger w \\ = ||A^\dagger z||_2^2 + (A^\dagger z)^\dagger w + w^\dagger (A^\dagger z) + ||w||_2^2 \\ = ||A^\dagger z||_2^2 + ||w||_2^2, \end{aligned}$$

since

$$w^\dagger (A^\dagger z) = (Aw)^\dagger z = 0^\dagger z = 0$$

and

$$(A^\dagger z)^\dagger w = z^\dagger A w = z^\dagger 0 = 0.$$

Therefore, $||x + w||_2 = ||A^\dagger z + w||_2 > ||A^\dagger z||_2 = ||x||_2$ unless $w = 0$. This completes the proof. ∎

In a later chapter we shall consider other approaches to solving underdetermined systems of linear equations.

3.10 Over-Determined Systems of Linear Equations

When there are more equations than there are unknowns in the system $Ax = b$ we say that the system is *over-determined*; it is most likely then that there will be no exact solution, although there may be (give an example). In such cases, it is common to seek a *least squares* solution. A least squares solution is not an exact solution of $Ax = b$ when none exist, but rather an exact solution of the system $A^\dagger Ax = A^\dagger b$. A least squares solution is a minimizer of the function

$$f(x) = \frac{1}{2}\|Ax - b\|_2^2.$$

Ex. 3.33 *Let A be an M by N matrix with complex entries. View A as a linear function with domain $\mathbb{C}^N$, the space of all N-dimensional complex column vectors, and range contained within $\mathbb{C}^M$, via the expression $A(x) = Ax$. Suppose that $M > N$. The range of A, denoted $R(A)$, cannot be all of $\mathbb{C}^M$. Show that every vector z in $\mathbb{C}^M$ can be written uniquely in the form $z = Ax + w$, where $A^\dagger w = 0$. Show that $\|z\|_2^2 = \|Ax\|_2^2 + \|w\|_2^2$, where $\|z\|_2^2$ denotes the square of the two-norm of z. Hint: If $z = Ax + w$ then consider $A^\dagger z$. Assume $A^\dagger A$ is invertible.*

Chapter 4

The ART, MART and EM-MART

4.1 Chapter Summary

The ART and the MART are two iterative algorithms that were designed to address issues that arose in solving large-scale systems of linear equations for medical imaging [151]. The EM-MART is a more recently discovered method that combines useful features of both ART and MART [54]. In this chapter we give an overview of these methods; later, we shall revisit them in more detail.

4.2 Overview

In many applications, such as in image processing, we need to solve a system of linear equations that is quite large, often several tens of thousands

of equations in about the same number of unknowns. In these cases, issues such as the costs of storage and retrieval of matrix entries, the computation involved in apparently trivial operations, such as matrix-vector products, and the speed of convergence of iterative methods demand greater attention. At the same time, the systems to be solved are often underdetermined, and solutions satisfying certain additional constraints, such as non-negativity, are required.

Both the *algebraic reconstruction technique* (ART) and the *multiplicative algebraic reconstruction technique* (MART) were introduced as two iterative methods for discrete image reconstruction in transmission tomography.

Both methods are what are called *row-action* methods, meaning that each step of the iteration uses only a single equation from the system. The MART is limited to non-negative systems for which non-negative solutions are sought. In the under-determined case, both algorithms find the solution closest to the starting vector, in the two-norm or weighted two-norm sense for ART, and in the cross-entropy sense for MART, so both algorithms can be viewed as solving optimization problems. In the appendix "Geometric Programming and the MART" we describe the use of MART to solve the dual geometric programming problem. For both algorithms, the starting vector can be chosen to incorporate prior information about the desired solution. In addition,the ART can be employed in several ways to obtain a least-squares solution, in the over-determined case.

The *simultaneous* MART (SMART) algorithm is a simultaneous variant of the MART in which all the equations are employed at each step of the iteration. Closely related to the SMART is the *expectation maximization maximum likelihood* (EMML) method, which is also a simultaneous algorithm.

The EM-MART is a row-action variant of the EMML algorithm. Like MART, it applies to non-negative systems of equations and produces non-negative solutions, but, like ART, does not require exponentiation, so is computationally simpler than MART.

4.3 The ART in Tomography

In x-ray transmission tomography, as an x-ray beam passes through the body, it encounters various types of matter, such as soft tissue, bone, ligaments, air, each weakening the beam to a greater or lesser extent. If the intensity of the beam upon entry is I_{in} and I_{out} is its lower intensity after

passing through the body, then, at least approximately,

$$I_{out} = I_{in} e^{-\int_L f},$$

where $f = f(x, y) \geq 0$ is the *attenuation function* describing the two-dimensional distribution of matter within the slice of the body being scanned and $\int_L f$ is the integral of the function f over the line L along which the x-ray beam has passed. This is the continuous model. In the discrete model the slice of the body being scanned is viewed as consisting of pixels, which we number $j = 1, 2, ..., J$. The x-rays are sent into the body along I lines, which we number $i = 1, 2, ..., I$. The line integral of f along the ith line is measured, approximately, from the entering and exiting strengths of the x-ray beams; these measurements are denoted b_i.

For $i = 1, ..., I$, let L_i be the set of pixel indices j for which the j-th pixel intersects the i-th line segment, as shown in Figure 4.1, and let $|L_i|$ be the cardinality of the set L_i. Let $A_{ij} = 1$ for j in L_i, and $A_{ij} = 0$ otherwise. With $i = k(\mathrm{mod}\, I) + 1$, the iterative step of the ART algorithm is

$$x_j^{k+1} = x_j^k + \frac{1}{|L_i|}(b_i - (Ax^k)_i), \tag{4.1}$$

for j in L_i, and

$$x_j^{k+1} = x_j^k, \tag{4.2}$$

if j is not in L_i. In each step of ART, we take the error, $b_i - (Ax^k)_i$, associated with the current x^k and the i-th equation, and distribute it equally over each of the pixels that intersects L_i.

This model is too simple; we are assuming that if the line segment intersects a pixel, then the entire amount of attenuating material within that pixel affects the x-ray strength. A somewhat more sophisticated version of ART allows A_{ij} to include the length of the i-th line segment that lies within the j-th pixel; A_{ij} is taken to be the ratio of this length to the length of the diagonal of the j-pixel.

More generally, ART can be viewed as an iterative method for solving an arbitrary system of linear equations, $Ax = b$.

4.4 The ART in the General Case

Let A be a matrix with complex entries, having I rows and J columns, and let b be a member of $\mathbb{C}^I$. We want to solve the system $Ax = b$. Note that when we say that A is a complex matrix and b a complex vector, we do not exclude the case in which the entries of both A and b are real.

Ex. 4.1 *Find the point in* $\mathbb{R}^2$ *on the line* $y = -3x + 6$ *closest to the point* $(4, 2)$.

Ex. 4.2 *Find the point in* $\mathbb{R}^3$ *on the plane* $x + 2y - 3z = 12$ *closest to the point* $(1, 1, 1)$.

Associated with each equation $(Ax)_i = b_i$ in the system $Ax = b$ there is a hyperplane H_i defined to be the subset of J-dimensional column vectors given by

$$H_i = \{x | (Ax)_i = b_i\}. \tag{4.3}$$

Ex. 4.3 *Show that the* i*th column of* $A^\dagger$ *is normal to the hyperplane* H_i*; that is, it is orthogonal to every vector lying in* H_i.

Ex. 4.4 *Show that, for any vector* z *in* $\mathbb{C}^J$*, the member of* H_i *closest to* z *is* x *having the entries*

$$x_j = z_j + \alpha_i^{-1}\overline{A_{ij}}(b_i - (Az)_i), \tag{4.4}$$

where

$$\alpha_i = \sum_{j=1}^{J} |A_{ij}|^2.$$

Definition 4.1 *The* orthogonal projection operator *onto the hyperplane* H_i *is the function* $P_i : \mathbb{C}^J \to \mathbb{C}^J$ *defined for each* z *in* $\mathbb{C}^J$ *by* $P_i z = x$*, where* x *is the member of* H_i *closest to* z.

The ART algorithm can be expressed in terms of the operators P_i. Let x^0 be arbitrary and, for each nonnegative integer k, let $i(k) = k(\mathrm{mod}\, I) + 1$. The iterative step of the ART is

$$x^{k+1} = P_{i(k)} x^k. \tag{4.5}$$

Using the formula in Equation (4.4), we can write the iterative step of the ART explicitly.

Algorithm 4.1 (ART) *For* $k = 0, 1, ...$ *and* $i = i(k) = k(\mathrm{mod}\, I) + 1$*, the entries of* x^{k+1} *are*

$$x_j^{k+1} = x_j^k + \alpha_i^{-1}\overline{A_{ij}}(b_i - (Ax^k)_i). \tag{4.6}$$

Because the ART uses only a single equation at each step, it has been called a *row-action* method.

4.4.1 Simplifying the Notation

To simplify our notation, we shall assume, throughout this chapter, that the rows of A have been rescaled to have Euclidean length one; that is

$$\alpha_i = \sum_{j=1}^{J} |A_{ij}|^2 = 1, \tag{4.7}$$

for each $i = 1, ..., I$, and that the entries of b have been rescaled accordingly, to preserve the equations $Ax = b$. The ART is then the following: begin with an arbitrary vector x^0; for each nonnegative integer k, having found x^k, the next iterate x^{k+1} has entries

$$x_j^{k+1} = x_j^k + \overline{A_{ij}}(b_i - (Ax^k)_i). \tag{4.8}$$

4.4.2 Consistency

When we are dealing with a general system of linear equations $Ax = b$, we shall say that the system is *consistent* if there are vectors x with $Ax = b$; that is, the system has exact solutions. If not, the system will be called *inconsistent*.

When we are dealing with non-negative systems $Ax = b$, in which the entries of A are non-negative, the entries of b are positive, and we seek a non-negative solution x, we shall say that such a system is consistent if there are non-negative vectors x with $Ax = b$; otherwise, the system is inconsistent. It will always be clear from the context which category of systems we are discussing. The ART applies to general systems of linear equations, while the MART and EM-MART apply only to non-negative systems. Note that a non-negative system can be inconsistent even when it possesses exact solutions x that happen not to be non-negative.

4.4.3 When $Ax = b$ Has Solutions

For the consistent case we have the following result concerning the ART.

Theorem 4.1 *Let $A\hat{x} = b$ and let x^0 be arbitrary. Let $\{x^k\}$ be generated by Equation (4.8). Then the sequence of Euclidean distances or two-norms $\{||\hat{x} - x^k||_2\}$ is decreasing and $\{x^k\}$ converges to the solution of $Ax = b$ closest to x^0.*

So, when the system $Ax = b$ has exact solutions, the ART converges to the solution closest to x^0, in the 2-norm. How fast the algorithm converges will depend on the ordering of the equations and on whether or not we use *relaxation*, which we shall discuss later. In selecting the equation ordering, the important thing is to avoid particularly bad orderings, in which the hyperplanes H_i and H_{i+1} are nearly parallel.

4.4.4 When $Ax = b$ Has No Solutions

When there are no exact solutions, the ART does not converge to a single vector, but, for each fixed i, the subsequence $\{x^{nI+i}, n = 0, 1, ...\}$ converges to a vector z^i and the collection $\{z^i \,|i = 1, ..., I\}$ is called the *limit cycle*. The ART limit cycle will vary with the ordering of the equations, and contains more than one vector unless an exact solution exists.

Figures 4.2 and 4.3 illustrate the behavior of the ART in the two cases.

4.4.5 The Geometric Least-Squares Solution

When the system $Ax = b$ has no solutions, it is reasonable to seek an approximate solution, such as the *least squares* solution, $x_{LS} = (A^\dagger A)^{-1}A^\dagger b$, which minimizes $||Ax - b||_2$. It is important to note that the system $Ax = b$ has solutions if and only if the related system $WAx = Wb$ has solutions, where W denotes an invertible matrix; when solutions of $Ax = b$ exist, they are identical to those of $WAx = Wb$. But, when $Ax = b$ does not have solutions, the least-squares solutions of $Ax = b$, which need not be unique, but usually are, and the least-squares solutions of $WAx = Wb$ need not be identical. In the typical case in which $A^\dagger A$ is invertible, the unique least-squares solution of $Ax = b$ is

$$(A^\dagger A)^{-1}A^\dagger b, \tag{4.9}$$

while the unique least-squares solution of $WAx = Wb$ is

$$(A^\dagger W^\dagger WA)^{-1}A^\dagger W^\dagger b, \tag{4.10}$$

and these need not be the same.

A simple example is the following. Consider the system

$$x = 1$$

$$x = 2, \tag{4.11}$$

which has the unique least-squares solution $x = 1.5$, and the system

$$2x = 2$$

$$x = 2, \tag{4.12}$$

which has the least-squares solution $x = 1.2$.

Definition 4.2 *The* geometric least-squares *solution of $Ax = b$ is the least-squares solution of $WAx = Wb$, for W the diagonal matrix whose entries are the reciprocals of the Euclidean lengths of the rows of A.*

In our example above, the geometric least-squares solution for the first system is found by using $W_{11} = 1 = W_{22}$, so is again $x = 1.5$, while the geometric least-squares solution of the second system is found by using $W_{11} = 0.5$ and $W_{22} = 1$, so that the geometric least-squares solution is $x = 1.5$, not $x = 1.2$.

Open Question: If there is a unique geometric least-squares solution, where is it, in relation to the vectors of the limit cycle? Can it be calculated easily, from the vectors of the limit cycle?

There is a partial answer to the first question. It is known that if the system $Ax = b$ has no exact solution, and if $I = J+1$, then the vectors of the limit cycle lie on a sphere in J-dimensional space having the geometric least-squares solution at its center [55]. This is not generally true for $I \neq J + 1$, however.

4.5 The MART

The *multiplicative* ART (MART) is an iterative algorithm closely related to the ART. It also was devised to obtain tomographic images, but, like ART, applies more generally; MART applies to non-negative systems of linear equations $Ax = b$ for which the b_i are positive, the A_{ij} are non-negative, and the solution x we seek is to have nonnegative entries. It is not so easy to see the relation between ART and MART if we look at the most general formulation of MART. For that reason, we begin with a simpler case, transmission tomographic imaging, in which the relation is most clearly apparent.

4.5.1 A Special Case of MART

We begin by considering the application of MART to the transmission tomography problem. For $i = 1, ..., I$, let L_i be the set of pixel indices j for which the j-th pixel intersects the i-th line segment, and let $|L_i|$ be the cardinality of the set L_i. Let $A_{ij} = 1$ for j in L_i, and $A_{ij} = 0$ otherwise. With $i = k(\mathrm{mod}\, I) + 1$, the iterative step of the ART algorithm is

$$x_j^{k+1} = x_j^k + \frac{1}{|L_i|}(b_i - (Ax^k)_i), \tag{4.13}$$

for j in L_i, and

$$x_j^{k+1} = x_j^k, \tag{4.14}$$

if j is not in L_i. In each step of ART, we take the error, $b_i - (Ax^k)_i$, associated with the current x^k and the i-th equation, and distribute it equally over each of the pixels that intersects L_i.

Suppose, now, that each b_i is positive, and we know in advance that the desired image we wish to reconstruct must be nonnegative. We can begin with $x^0 > 0$, but as we compute the ART steps, we may lose nonnegativity. One way to avoid this loss is to correct the current x^k multiplicatively, rather than additively, as in ART. This leads to the *multiplicative* ART (MART).

The MART, in this case, has the iterative step

$$x_j^{k+1} = x_j^k \Big(\frac{b_i}{(Ax^k)_i} \Big), \tag{4.15}$$

for those j in L_i, and

$$x_j^{k+1} = x_j^k, \tag{4.16}$$

otherwise. Therefore, we can write the iterative step as

$$x_j^{k+1} = x_j^k \Big(\frac{b_i}{(Ax^k)_i} \Big)^{A_{ij}}. \tag{4.17}$$

4.5.2 The MART in the General Case

Taking the entries of the matrix A to be either one or zero, depending on whether or not the j-th pixel is in the set L_i, is too crude. The line L_i may just clip a corner of one pixel, but pass through the center of another. Surely, it makes more sense to let A_{ij} be the length of the intersection of line L_i with the j-th pixel, or, perhaps, this length divided by the length of the diagonal of the pixel. It may also be more realistic to consider a strip, instead of a line. Other modifications to A_{ij} may be made, in order to better describe the physics of the situation. Finally, all we can be sure of is that A_{ij} will be nonnegative, for each i and j. In such cases, what is the proper form for the MART?

The MART, which can be applied only to nonnegative systems, is a sequential, or row-action, method that uses one equation only at each step of the iteration.

Algorithm 4.2 (MART) *Let x^0 be a positive vector. For $k = 0, 1, ...$, and $i = k(\mathrm{mod}\, I) + 1$, having found x^k define x^{k+1} by*

$$x_j^{k+1} = x_j^k \Big(\frac{b_i}{(Ax^k)_i} \Big)^{m_i^{-1} A_{ij}}, \tag{4.18}$$

where $m_i = \max \{A_{ij} \,|j = 1, 2, ..., J\}$.

Some treatments of MART leave out the m_i, but require only that the entries of A have been rescaled so that $A_{ij} \leq 1$ for all i and j. The m_i is important, however, in accelerating the convergence of MART.

Notice that we can write x_j^{k+1} as a weighted geometric mean of x_j^k and $x_j^k\left(\frac{b_i}{(Ax^k)_i}\right)$:

$$x_j^{k+1} = \left(x_j^k\right)^{1-m_i^{-1}A_{ij}}\left(x_j^k\left(\frac{b_i}{(Ax^k)_i}\right)\right)^{m_i^{-1}A_{ij}}. \tag{4.19}$$

This will help to motivate the EM-MART.

4.5.3 Cross-Entropy

For $a > 0$ and $b > 0$, let the cross-entropy or Kullback-Leibler (KL) distance from a to b be

$$KL(a, b) = a \log \frac{a}{b} + b - a, \tag{4.20}$$

with $KL(a, 0) = +\infty$, and $KL(0, b) = b$. Extend to nonnegative vectors coordinate-wise, so that

$$KL(x, z) = \sum_{j=1}^{J} KL(x_j, z_j). \tag{4.21}$$

Unlike the Euclidean distance, the KL distance is not symmetric; $KL(Ax, b)$ and $KL(b, Ax)$ are distinct, and we can obtain different approximate solutions of $Ax = b$ by minimizing these two distances with respect to non-negative x.

4.5.4 Convergence of MART

In the consistent case, by which we mean that $Ax = b$ has nonnegative solutions, we have the following convergence theorem for MART.

Theorem 4.2 *In the consistent case, the MART converges to the unique nonnegative solution of $b = Ax$ for which the distance $\sum_{j=1}^{J} KL(x_j, x_j^0)$ is minimized.*

If the starting vector x^0 is the vector whose entries are all one, then the MART converges to the solution that maximizes the *Shannon entropy*,

$$SE(x) = \sum_{j=1}^{J} x_j \log x_j - x_j. \tag{4.22}$$

As with ART, the speed of convergence is greatly affected by the ordering of the equations, converging most slowly when consecutive equations correspond to nearly parallel hyperplanes.

Open Question: When there are no nonnegative solutions, MART does not converge to a single vector, but, like ART, is always observed to produce a limit cycle of vectors. Unlike ART, there is no proof of the existence of a limit cycle for MART. Is there such a proof?

4.6 The EM-MART

The MART enforces positivity of the x_j^k, but at the cost of an exponentiation in each step. The EM-MART is similar to the MART, guarantees positivity at each step, but does not employ exponentiation.

The EM-MART is a row-action version of the *expectation maximization maximum likelihood* (EMML) algorithm, which we shall discuss in detail later. The EMML algorithm, which was developed as a method for reconstructing tomographic medical images, was found to converge too slowly to be of practical use. Several faster variants of the EMML algorithm were subsequently discovered, one of which is the EM-MART.

As with MART, we assume that the entries of the matrix A are nonnegative, that the entries of b are positive, and that we seek a non-negative solution of $Ax = b$.

Algorithm 4.3 (EM-MART) *Let x^0 be an arbitrary positive vector and $i = k(\mathrm{mod}\, I) + 1$. Then let*

$$x_j^{k+1} = (1 - m_i^{-1} A_{ij}) x_j^k + m_i^{-1} A_{ij} \Big(x_j^k \frac{b_i}{(Ax^k)_i} \Big). \tag{4.23}$$

Notice that x_j^{k+1} is always positive, since it is a weighted arithmetic mean of x_j^k and $x_j^k \Big(\frac{b_i}{(Ax^k)_i} \Big)$.

In the consistent case, in which there are non-negative solutions of $Ax = b$, the EM-MART converges to a non-negative solution. However, no characterization of the solution, in terms of x^0, is known.

The EM-MART fails to converge in the inconsistent case. What is always observed, but for which no proof is known, is that, for each fixed $i = 1, 2, ..., I$, as $m \to +\infty$, the EM-MART subsequences $\{x^{mI+i}\}$ converge to separate limit vectors, say $x^{\infty,i}$.

Open Questions: We know that, in the consistent case, the MART con-

verges to the non-negative solution of $Ax = b$ for which $KL(x, x^0)$ is minimized. Is there a similar characterization of the EM-MART solution, in terms of x^0? When there are no nonnegative solutions, EM-MART does not converge to a single vector, but, like ART and MART, is always observed to produce a limit cycle of vectors. Unlike ART, no one has found a proof of the existence of a limit cycle for EM-MART. Is there such a proof?

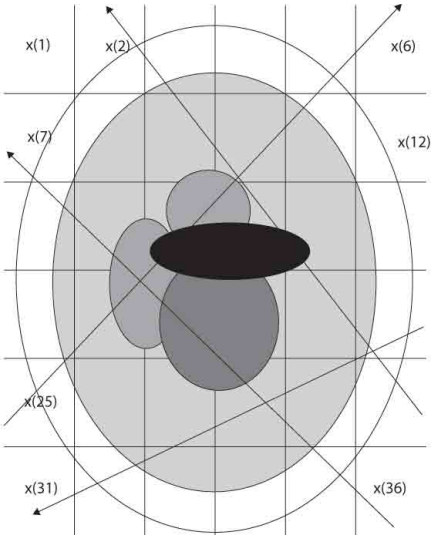

FIGURE 4.1: Line integrals through a discretized object.

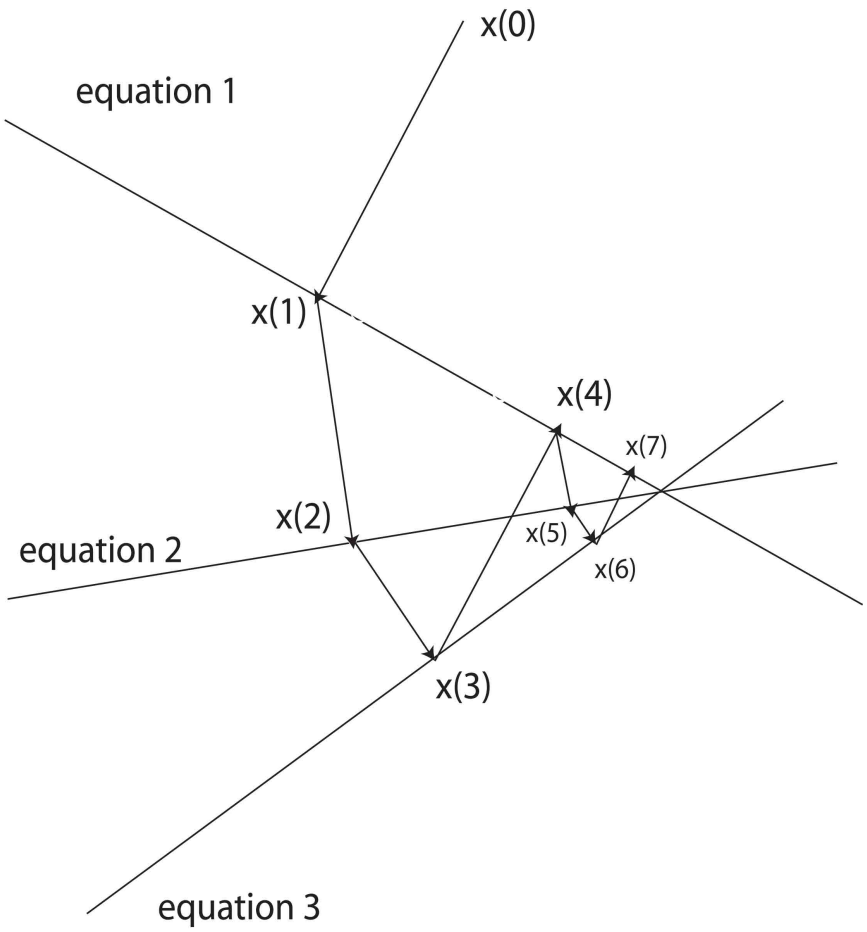

FIGURE 4.2: The ART algorithm in the consistent case.

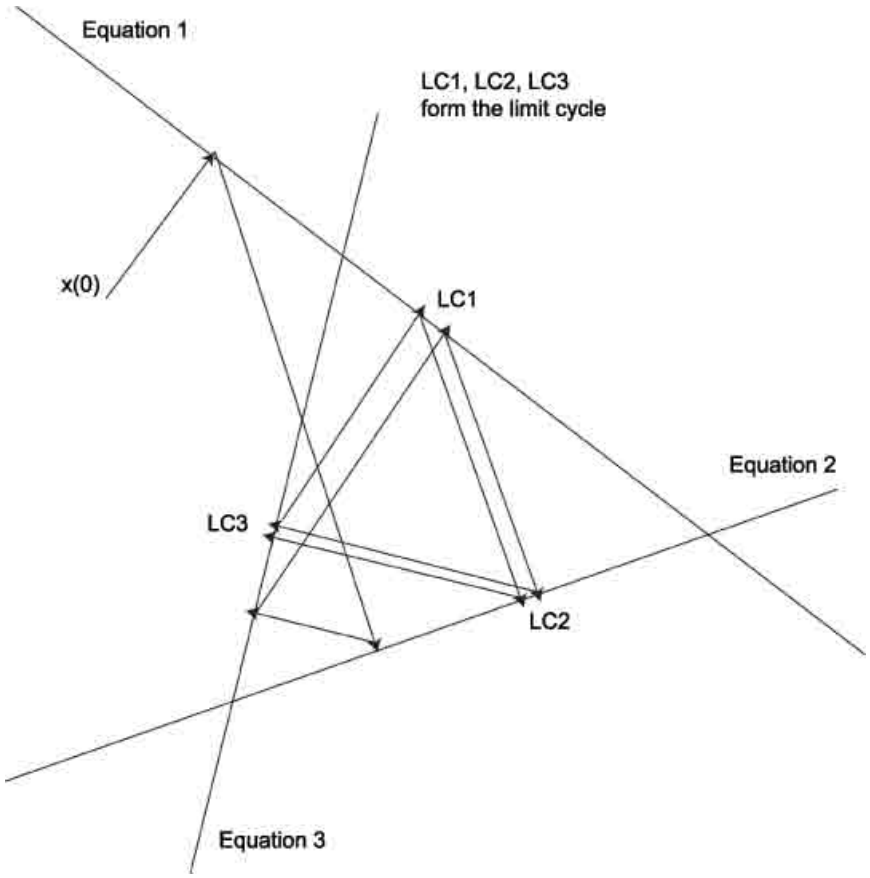

FIGURE 4.3: The ART algorithm in the inconsistent case.

Part II

Algebra

Chapter 5

More Matrix Theory

5.1 Chapter Summary

In this chapter we continue our study of matrix algebra.

5.2 Proof By Induction

Proof by induction is a tool used in a wide variety of proofs; we shall use it shortly to prove Schur's Lemma. In this section we present the basic idea and an example to illustrate its use.

All proofs by induction have the same basic form. There is some property, say Property P, that a positive integer n may or may not have. The assertion, which we must prove, is that all n have Property P. The proof is by contradiction; we assume the assertion is false and that not all n have Property P. Therefore, there must be a first n that does not have Property P. We begin by checking to see if $n = 1$ has Property P. Having established that $n = 1$ has Property P, we focus on the first n that does not have Property P; we know that this n is not one, so $n - 1$ is also a positive integer, and $n - 1$ does have Property P, since n is the first one without Property P. The rest of the proof involves showing that, because $n - 1$ has Property P, so must n. This will give us our contradiction and allow us to conclude that there is no such first n without Property P.

For example, let Property P be the following: n is a positive integer such that the sum of the first n positive integers is $\frac{1}{2}n(n+1)$. This clearly holds for $n = 1$, so $n = 1$ has Property P. Assume that not all n do have Property P, and let n be the first that does not have Property P. Then $n - 1$ is a positive integer and

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

Then

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

Therefore, n must also have Property P. This contradicts our assumption that not all positive integers have Property P. Therefore, Property P holds for all positive integers.

Note that there are other ways to prove this theorem. We have used induction here because we are trying to illustrate the use of induction. In most cases in which induction is used, induction is the best, and maybe the only, way to prove the theorem.

Ex. 5.1 *Prove that*

$$\frac{1}{2!} + \frac{2}{3!} + ... + \frac{n}{(n+1)!} = 1 - \frac{1}{(n+1)!}.$$

5.3 Schur's Lemma

Schur's Lemma is a useful tool for proving the diagonalization theorems for Hermitian and normal matrices.

Definition 5.1 *A complex square matrix U is said to be* unitary *if $U^\dagger U = I$. A real square matrix O is* orthogonal *if $O^T O = I$. A square matrix T is* upper triangular *if all the entries of T below the main diagonal are zero.*

Theorem 5.1 (Schur's Lemma) *For any square matrix S there is a unitary matrix U such that $U^\dagger S U = T$ is an upper triangular matrix.*

Proof: We proceed by induction. The theorem is obviously true for any 1 by 1 matrix. Assume that the theorem is true for any $n-1$ by $n-1$ matrix. We show that it is true also for any n by n matrix.

Because every polynomial has at least one (possibly complex) root, S has at least one eigenvector. Therefore, let $Su^1 = \lambda u^1$, with $\|u^1\|_2 = 1$. Let $\{u^1, u^2, ..., u^n\}$ be an orthonormal basis for $\mathbb{C}^n$. Then

$$U = \begin{bmatrix} u^1 & u^2 & ... & u^n \end{bmatrix} \tag{5.1}$$

is unitary and

$$U^\dagger S U = \begin{bmatrix} (u^1)^\dagger \\ (u^2)^\dagger \\ . \\ . \\ . \\ (u^n)^\dagger \end{bmatrix} \begin{bmatrix} Su^1 & Su^2 & ... & Su^n \end{bmatrix} = \begin{bmatrix} \lambda_1 & c_{12} & ... & c_{1n} \\ 0 & & & \\ 0 & & & \\ . & & S_1 & \\ . & & & \\ . & & & \\ 0 & & & \end{bmatrix},$$

where S_1 is of order $n-1$.

Now let U_1 be an $n-1$ by $n-1$ unitary matrix such that $U_1^\dagger S_1 U_1$ is upper triangular; such a U_1 exists by the induction hypothesis. Let

$$U_2 = \begin{bmatrix} 1 & 0 & ... & 0 \\ 0 & & & \\ 0 & & & \\ . & & U_1 & \\ . & & & \\ . & & & \\ 0 & & & \end{bmatrix}.$$

Then U_2 and UU_2 are unitary and

$$(UU_2)^\dagger S(UU_2) = U_2^\dagger (U^\dagger SU)U_2$$

$$= \begin{bmatrix} \lambda_1 & b_{12} & \ldots & b_{1n} \\ 0 & & & \\ 0 & & & \\ . & & U_1^\dagger SU_1 & \\ . & & & \\ . & & & \\ 0 & & & \end{bmatrix},$$

which is upper triangular. ∎

Using essentially the same proof, we can establish the following version of Schur's Lemma:

Theorem 5.2 *Let S be a real square matrix with only real eigenvalues. Then there is a real orthogonal matrix O such that $O^T SO$ is upper triangular.*

Corollary 5.1

- *(a) If $S^\dagger = S$ then there is a unitary matrix U such that $U^\dagger SU$ is a real diagonal matrix.*
- *(b) If S is real and $S^T = S$ then there is an orthogonal matrix O such that $O^T SO$ is a real diagonal matrix.*

Ex. 5.2 *Use Schur's Lemma to prove Corollary 5.1.*

Theorem 5.3 *For a given complex square matrix S there is a unitary matrix U such that $U^\dagger SU = D$ is a diagonal matrix if and only if S is normal.*

Ex. 5.3 *Use Schur's Lemma to prove Theorem 5.3. Hint: compute $(TT^\dagger)_{nn}$ two ways.*

We are now in a position to prove Proposition 3.8, which we restate now.

Proposition 5.1 *A square matrix A is invertible if and only if its determinant is not zero.*

Proof: From Schur's Lemma we know that there is a unitary matrix U such that $U^\dagger AU = T$ is upper triangular. The determinant of T is the product of the entries on its main diagonal. Clearly, T is invertible if and only if none of these entries is zero; this is true because $Tx = 0$ implies $x = 0$ if and only if no diagonal entry is zero. Therefore, T is invertible if and only if the determinant of T is not zero. But, the determinant of A is the same as that of T and A is invertible precisely when T is invertible. ∎

5.4 Eigenvalues and Eigenvectors

Let A be a complex M by N matrix. It is often helpful to know how large the two-norm $\|Ax\|_2$ can be, relative to $\|x\|_2$; that is, we want to find a constant a so that

$$\|Ax\|_2/\|x\|_2 \leq a,$$

for all $x \neq 0$. We can reformulate the problem by asking how large $\|Au\|_2^2$ can be, subject to $\|u\|_2 = 1$. Using Lagrange multipliers, we discover that a unit vector u that maximizes $\|Au\|_2^2$ has the property that

$$A^\dagger Au = \lambda u,$$

for some constant λ. This leads to the more general problem discussed in this section.

Definition 5.2 *Given an N by N complex matrix S, we say that a complex number λ is an* eigenvalue *of S if there is a nonzero vector u with $Su = \lambda u$. The column vector u is then called an* eigenvector *of S associated with eigenvalue λ.*

Clearly, if u is an eigenvector of S, then so is cu, for any constant $c \neq 0$; therefore, it is common to choose eigenvectors to have norm equal to one.

If λ is an eigenvalue of S, then the matrix $S - \lambda I$ fails to have an inverse, since $(S - \lambda I)u = 0$ but $u \neq 0$, and so its determinant must be zero. If we treat λ as a variable and compute the *characteristic polynomial* of S,

$$P(\lambda) = \det(S - \lambda I),$$

we obtain a polynomial of degree N in λ. Its roots $\lambda_1, ..., \lambda_N$ are then the eigenvalues of S. If $||u||_2^2 = u^\dagger u = 1$ then $u^\dagger Su = \lambda u^\dagger u = \lambda$. Note that the eigenvalues need not be real, even if S is a real matrix.

Ex. 5.4 *Prove that the eigenvalues of an upper triangular matrix T are the entries of its main diagonal, so that the trace of T is the sum of its eigenvalues.*

Ex. 5.5 *Prove that, if S is square, U is unitary, and $U^\dagger SU = T$ is upper triangular, then the eigenvalues of S and T are the same and S and T have the same trace. Hint: use the facts that* $\det(AB) = \det(A)\det(B)$ *and Equation (3.22).*

Ex. 5.6 *Use the two previous exercises to prove that, for any square matrix S, the trace of S is the sum of its eigenvalues.*

We know that a square matrix S is invertible if and only if $Sx = 0$ implies that $x = 0$. We can say this another way now: S is invertible if and only if $\lambda = 0$ is not an eigenvalue of S.

Ex. 5.7 *Compute the eigenvalues for the real square matrix*

$$S = \begin{bmatrix} 1 & 2 \\ -2 & 1 \end{bmatrix}. \tag{5.2}$$

Note that the eigenvalues are complex, even though the entries of S are real.

The eigenvalues of the Hermitian matrix

$$H = \begin{bmatrix} 1 & 2+i \\ 2-i & 1 \end{bmatrix} \tag{5.3}$$

are $\lambda = 1 + \sqrt{5}$ and $\lambda = 1 - \sqrt{5}$, with corresponding eigenvectors $u = (\sqrt{5}, 2 - i)^T$ and $v = (\sqrt{5}, i - 2)^T$, respectively. Then $\tilde{H}$, defined as in Equation (3.31), has the same eigenvalues, but both with multiplicity two. Finally, the associated eigenvectors of $\tilde{B}$ are

$$\begin{bmatrix} u^1 \\ u^2 \end{bmatrix}, \tag{5.4}$$

and

$$\begin{bmatrix} -u^2 \\ u^1 \end{bmatrix}, \tag{5.5}$$

for $\lambda = 1 + \sqrt{5}$, and

$$\begin{bmatrix} v^1 \\ v^2 \end{bmatrix}, \tag{5.6}$$

and

$$\begin{bmatrix} -v^2 \\ v^1 \end{bmatrix}, \tag{5.7}$$

for $\lambda = 1 - \sqrt{5}$.

Definition 5.3 *The* spectral radius *of S, denoted $\rho(S)$, is the largest value of $|\lambda|$, where λ denotes an eigenvalue of S.*

Ex. 5.8 *Use the facts that λ is an eigenvalue of S if and only if $det(S - \lambda I) = 0$, and $det(AB) = det(A)det(B)$ to show that λ^2 is an eigenvalue of S^2 if and only if either λ or $-\lambda$ is an eigenvalue of S. Then use this result to show that $\rho(S)^2 = \rho(S^2)$.*

5.4.1 The Hermitian Case

Let H be an N by N Hermitian matrix. As we just saw, there is a unitary matrix U such that $U^\dagger HU = D$ is real and diagonal. Then $HU = UD$, so that the columns of U are eigenvectors of H with two-norms equal to one, and the diagonal entries of D are the eigenvalues of H. Since U is invertible, its columns form a set of N mutually orthogonal norm-one eigenvectors of the Hermitian matrix H; call them $\{u^1, ..., u^N\}$. We denote by λ_n, $n = 1, 2, ..., N$, the N eigenvalues, so that $Hu^n = \lambda_n u^n$. This is the well known *eigenvalue-eigenvector decomposition* of the matrix H. Not every square matrix has such a decomposition, which is why we focus on Hermitian H. The singular-value decomposition, which we discuss shortly, provides a similar decomposition for an arbitrary, possibly non-square, matrix.

The matrix H can also be written as

$$H = \sum_{n=1}^{N} \lambda_n u^n (u^n)^\dagger,$$

a linear superposition of the *dyad* matrices $u^n(u^n)^\dagger$. The Hermitian matrix H is invertible if and only if none of the λ are zero and its inverse is

$$H^{-1} = \sum_{n=1}^{N} \lambda_n^{-1} u^n (u^n)^\dagger.$$

We also have $H^{-1} = UL^{-1}U^\dagger$.

Ex. 5.9 *Show that if $z = (z_1, ..., z_N)^T$ is a column vector with complex entries and $H = H^\dagger$ is an N by N Hermitian matrix with complex entries then the quadratic form $z^\dagger Hz$ is a real number. Show that the quadratic form $z^\dagger Hz$ can be calculated using only real numbers. Let $z = x + iy$, with x and y real vectors and let $H = A + iB$, where A and B are real matrices. Then show that $A^T = A$, $B^T = -B$, $x^T Bx = 0$ and finally,*

$$z^\dagger Hz = \begin{bmatrix} x^T & y^T \end{bmatrix} \begin{bmatrix} A & -B \\ B & A \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}.$$

Use the fact that $z^\dagger Hz$ is real for every vector z to conclude that the eigenvalues of H are real.

Ex. 5.10 *Show that the eigenvalues of a Hermitian matrix H are real by computing the conjugate transpose of the 1 by 1 matrix $z^\dagger Hz$.*

Definition 5.4 *A Hermitian matrix Q is said to be* nonnegative-definite *if all the eigenvalues of Q are nonnegative, and* positive-definite *if all the eigenvalues are positive.*

Proposition 5.2 *A Hermitian matrix Q is a nonnegative-definite matrix if and only if there is another matrix C, not necessarily square, such that $Q = C^\dagger C$.*

Proof: Assume that Q is nonnegative-definite and let $Q = ULU^\dagger$ be the eigenvalue/eigenvector decomposition of Q. Since the eigenvalues of Q are nonnegative, each diagonal entry of the matrix L has a nonnegative square root; the matrix with these square roots as entries is called $\sqrt{L}$. Using the fact that $U^\dagger U = I$, we have

$$Q = ULU^\dagger = U\sqrt{L}U^\dagger U\sqrt{L}U^\dagger;$$

we then take $C = U\sqrt{L}U^\dagger$, so $C^\dagger = C$. This choice of C is called the *Hermitian square root* of Q.

Conversely, assume now that $Q = C^\dagger C$, for some arbitrary, possibly not square, matrix C. Let $Qu = \lambda u$, for some non-zero eigenvector u, so that λ is an eigenvalue of Q. Then

$$\lambda\|u\|_2^2 = \lambda u^\dagger u = u^\dagger Q u = u^\dagger C^\dagger C u = \|Cu\|_2^2,$$

so that

$$\lambda = \|Cu\|_2^2/\|u\|_2^2 \geq 0.$$

∎

If N is a square complex matrix with $N = UDU^\dagger$, where, as above, $U^\dagger U = I$ and D is diagonal, but not necessarily real, then we do have $N^\dagger N = NN^\dagger$; then N is *normal*, which means that $N^T N = NN^T$. The matrix N will be Hermitian if and only if D is real. It follows then that a real normal matrix N will be symmetric if and only if its eigenvalues are real, since it is then Hermitian and real.

The normal matrices are precisely those for which such an eigenvector-eigenvalue decomposition holds, as we saw above. In the appendix on Hermitian and Normal Linear Operators we prove this result again, as a statement about operators on a finite-dimensional vector space.

The following exercise gives an example of a matrix N that is real, normal, not symmetric, and has non-real eigenvalues. The matrix $N^T N$ has repeated eigenvalues. As we shall see in Theorem 5.4, if a real, normal matrix is such that $N^T N$ does not have repeated eigenvalues, then N is symmetric and so the eigenvalues of N are real.

Ex. 5.11 *Show that the 2 by 2 matrix $N = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$ is real, normal, and has eigenvalues $\pm i$. Show that the eigenvalues of $N^T N$ are both 1.*

5.5 The Singular Value Decomposition (SVD)

The year 1965 was a good one for the discovery of important algorithms. In that year, Cooley and Tukey [101] introduced the *fast Fourier transform* (FFT) algorithm and Golub and Kahan [149] their method for calculating the *singular-value decomposition* (SVD).

We have just seen that an N by N Hermitian matrix H can be written in terms of its eigenvalues and eigenvectors as $H = ULU^\dagger$ or as

$$H = \sum_{n=1}^{N} \lambda_n u^n (u^n)^\dagger.$$

The *singular value decomposition* (SVD) is a similar result that applies to any rectangular matrix A. It is an important tool in image compression and pseudo-inversion.

5.5.1 Defining the SVD

Let A be any M by N complex matrix. In presenting the SVD of A we shall assume that $N \geq M$; the SVD of $A^\dagger$ will come from that of A. Let $Q = A^\dagger A$ and $P = AA^\dagger$; we assume, reasonably, that P, the smaller of the two matrices, is invertible, so all the eigenvalues $\lambda_1, ..., \lambda_M$ of P are positive. We let the eigenvalue/eigenvector decomposition of P be $P = ULU^\dagger$, where $\{u^1, ..., u^M\}$ are orthonormal eigenvectors of P and $Pu^m = \lambda_m u^m$.

From $PU = UL$ or $AA^\dagger U = UL$ it follows that $A^\dagger AA^\dagger U = A^\dagger UL$. Therefore, the M columns of $W = A^\dagger U$ are eigenvectors of Q corresponding to the eigenvalues λ_m; since $Pu^m = AA^\dagger u^m$ is not the zero vector, $A^\dagger u^m$ cannot be the zero vector either. But the columns of W do not have norm one. To normalize these columns we replace them with the M columns of $A^\dagger UL^{-1/2}$, which are orthonormal eigenvectors of Q.

Ex. 5.12 *Show that the nonzero eigenvalues of $Q = A^\dagger A$ and $P = AA^\dagger$ are the same.*

Let Z be the N by N matrix whose first M columns are those of the matrix $A^\dagger UL^{-1/2}$ and whose remaining $N - M$ columns are any mutually orthogonal norm-one vectors that are all orthogonal to each of the first M columns; note that this gives us $Z^\dagger Z = I$.

Let Σ be the M by N matrix with diagonal entries $\Sigma_{mm} = \sqrt{\lambda_m}$, for $m = 1, ..., M$, and whose remaining entries are zero. The nonzero entries of Σ, the $\sqrt{\lambda_m}$, are called the *singular values* of A. The *singular value decomposition* (SVD) of A is $A = U\Sigma Z^\dagger$. The SVD of $A^\dagger$ is $A^\dagger = Z\Sigma^T U^\dagger$.

Ex. 5.13 *Show that $U\Sigma Z^\dagger$ equals A.*

We have assumed, for convenience, that none of the eigenvalues λ_m, $m = 1, ..., M$ are zero. If this is not true, we can obtain the SVD of A simply by modifying the definition of $L^{-1/2}$ to have $1/\sqrt{\lambda_m}$ on the main diagonal if λ_m is not zero, and zero if it is. To show that $U\Sigma Z^\dagger = A$ now we need to use the fact that $Pu^m = 0$ implies that $A^\dagger u^m = 0$. To see this, note that

$$0 = Pu^m = AA^\dagger u^m$$

implies that

$$0 = (u^m)^\dagger Pu^m = (u^m)^\dagger AA^\dagger u^m = \|A^\dagger u^m\|^2.$$

As an example of the singular-value decomposition, consider the matrix A, whose SVD is given by

$$A = \begin{bmatrix} 4 & 8 & 8 \\ 3 & 6 & 6 \end{bmatrix} = \begin{bmatrix} 4/5 & 3/5 \\ 3/5 & -4/5 \end{bmatrix} \begin{bmatrix} 15 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1/3 & 2/3 & 2/3 \\ 2/3 & -2/3 & 1/3 \\ 2/3 & 1/3 & -2/3 \end{bmatrix},$$

which can also be written in dyad form as

$$A = 15 \begin{bmatrix} 4/5 \\ 3/5 \end{bmatrix} \begin{bmatrix} 1/3 & 2/3 & 2/3 \end{bmatrix}.$$

It is just a coincidence that, in this example, the matrices U and Z are symmetric.

The SVD of A^T is then

$$A^T = \begin{bmatrix} 4 & 3 \\ 8 & 6 \\ 8 & 6 \end{bmatrix} = \begin{bmatrix} 1/3 & 2/3 & 2/3 \\ 2/3 & -2/3 & 1/3 \\ 2/3 & 1/3 & -2/3 \end{bmatrix} \begin{bmatrix} 15 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 4/5 & 3/5 \\ 3/5 & -4/5 \end{bmatrix}.$$

Ex. 5.14 *If H is a Hermitian matrix, its eigenvalue/eigenvector decomposition $H = ULU^\dagger$ need not be its SVD. Illustrate this point for the real symmetric matrix* $\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$.

Using the SVD of A we can write A as a sum of dyads:

$$A = \sum_{m=1}^{M} \sqrt{\lambda_m} u^m (z^m)^\dagger, \tag{5.8}$$

where z^m denotes the mth column of the matrix Z.

In image processing, matrices such as A are used to represent discrete two-dimensional images, with the entries of A corresponding to the grey level or color at each pixel. It is common to find that most of the M singular values of A are nearly zero, so that A can be written approximately as a sum of far fewer than M dyads; this leads to SVD image compression. Such compression is helpful when many images are being transmitted, as, for example, when pictures of the surface of Mars are sent back to Earth.

Figures 5.1 and 5.2 illustrate what can be achieved with SVD compression. In both Figures the original is in the upper left. It is a 128 by 128 digitized image, so $M = 128$. In the images that follow, the number of terms retained in the sum in Equation (5.8) is, first, 2, then 4, 6, 8, 10, 20 and finally 30. The full sum has 128 terms, remember. In Figure 5.1 the text is nearly readable using only 10 terms, and certainly could be made perfectly readable with suitable software, so storing just this compressed image would be acceptable. In Figure 5.2, an image of a satellite, we get a fairly good idea of the general shape of the object from the beginning, with only two terms.

Ex. 5.15 *Suppose that $M = N$ and A is invertible. Show that we can write*

$$A^{-1} = \sum_{m=1}^{M} (\sqrt{\lambda_m})^{-1} z^m (u^m)^\dagger.$$

5.5.2 An Application in Space Exploration

The *Galileo* was deployed from the space shuttle *Atlantis* on October 18, 1989. After a detour around Venus and back past Earth to pick up gravity-assisted speed, *Galileo* headed for Jupiter. Its mission included a study of Jupiter's moon Europa, and the plan was to send back one high-resolution photo per minute, at a rate of $134KB$ per second, via a huge high-gain antenna. When the time came to open the antenna, it stuck. Without the pictures, the mission would be a failure.

There was a much smaller *low-gain* antenna on board, but the best transmission rate was going to be ten bits per second. All that could be done from earth was to reprogram an old on-board computer to compress the pictures prior to transmission. The problem was that pictures could be taken much faster than they could be transmitted to earth; some way to store them prior to transmission was key. The original designers of the software had long since retired, but the engineers figured out a way to introduce state-of-the art image compression algorithms into the computer. It happened that there was an ancient reel-to-reel storage device on board that was there only to serve as a backup for storing atmospheric data.

Using this device and the compression methods, the engineers saved the mission [16].

5.5.3 A Theorem on Real Normal Matrices

Consider the real square matrix

$$S = \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}.$$

Since

$$S^T S = S S^T = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix},$$

S is a real normal matrix. The eigenvalues of S are complex, S is not symmetric, and the eigenvalues of $S^T S$ are not distinct. In contrast, we have the following theorem.

Let N be a real square matrix that is normal; that is $N^T N = N N^T$. Now we use the SVD of N to prove the following theorem.

Theorem 5.4 *If N is a real normal matrix and all the eigenvalues of $N^T N$ are distinct, then N is symmetric.*

Proof: Let $Q = N^T N$. Since Q is real, symmetric, and non-negative definite, there is an orthogonal matrix O such that $QO = NN^T O = OD^2$, with $D \geq 0$ and D^2 the diagonal matrix whose diagonal entries are the eigenvalues of $Q = N^T N$. We shall want to be able to assume that the entries of D are all positive, which requires a bit of explanation.

We replace the matrix N with the new matrix $N + \alpha I$, where $\alpha > 0$ is selected so that the matrix $(N+\alpha I)(N+\alpha I)^T$ has only positive eigenvalues. We can do this because

$$(N + \alpha I)(N + \alpha I)^T = NN^T + \alpha(N + N^T) + \alpha^2 I;$$

the first and third matrices have only non-negative eigenvalues and the second one has only real ones, so a large enough α can be found. Now we can prove the theorem for the new matrix $N + \alpha I$, showing that it is symmetric. But it then follows that the matrix N must also be symmetric.

Now we continue with the proof, assuming that $D > 0$. The columns of $Z = N^T O D^{-1}$ are then orthonormal eigenvectors of $N^T N$ and the SVD of N is $N = ODZ^T$.

Since N is normal, we have $N^T NO = OD^2$, and

$$ZD^2 = N^T NZ = OD^2 O^T Z,$$

so that

$$O^T ZD^2 = D^2 O^T Z.$$

It follows from Exercise 3.7 that $O^T Z = B$ is diagonal. From $Z = OB$ and

$$N = ODZ^T = ODB^T O^T = ODBO^T = OCO^T,$$

where $C = DB$ is diagonal, it follows that $N^T = N$. ∎

This proof illustrates a use of the SVD of N, but the theorem can be proved using the eigenvector diagonalization of the normal matrix N itself. Note that the characteristic polynomial of N has real coefficients, so its roots occur in conjugate pairs. If N has a complex root λ, then both λ and $\overline{\lambda}$ are eigenvalues of N. It follows that $|\lambda|^2$ is an eigenvalue of $N^T N$ with multiplicity at least two. Consequently, if $N^T N$ has no repeated eigenvalues, then every eigenvalue of N is real. Using $U^\dagger N U = D$, with D real and diagonal, we get $N = UDU^\dagger$, so that $N^\dagger = UDU^\dagger = N$. Therefore N is real and Hermitian, and so is symmetric.

5.5.4 The Golub-Kahan Algorithm

We have obtained the SVD of A using the eigenvectors and eigenvalues of the Hermitian matrices $Q = A^\dagger A$ and $P = AA^\dagger$; for large matrices, this is not an efficient way to get the SVD. The Golub-Kahan algorithm [149] calculates the SVD of A without forming the matrices P and Q.

A matrix A is *bi-diagonal* if the only non-zero entries occur on the main diagonal and the first diagonal above the main one. Any matrix can be reduced to bi-diagonal form by multiplying the matrix first on the left by a succession of Householder matrices, and then on the right by another succession of Householder matrices. The QR factorization is easier to calculate when the matrix involved is bi-diagonal.

The Golub-Kahan algorithm for calculating the SVD of A involves first reducing A to a matrix B in bi-diagonal form and then applying a variant of the QR factorization.

Using Householder matrices, we get unitary matrices U_0 and Z_0 such that $A = U_0 B Z_0^\dagger$, where B is bi-diagonal. Then we find the SVD of B,

$$B = \tilde{U}\Sigma\tilde{Z}^\dagger,$$

using QR factorization. Finally, the SVD for A itself is

$$A = U_0\tilde{U}\Sigma\tilde{Z}^\dagger Z_0^\dagger.$$

Ever since the publication of the Golub-Kahan algorithm, there have been efforts to improve both the accuracy and the speed of the method. The improvements announced in [118] and [119] won for their authors the 2009 SIAM Activity Group on Linear Algebra Prize.

5.6 Generalized Inverses

Even if A does not have an inverse, as, for example, when $M \neq N$, it does have *generalized inverses* or *pseudo-inverses.*

Definition 5.5 *A matrix G is called a* generalized inverse *or* pseudo-inverse *for a matrix A if $x = Gb$ is a solution of $Ax = b$, whenever there are solutions.*

It is not obvious that generalized inverses exist for an arbitrary matrix A, but they do. In fact, we can use the SVD to obtain a pseudo-inverse for any A.

5.6.1 The Moore-Penrose Pseudo-Inverse

The *Moore-Penrose pseudo-inverse* is the matrix

$$A^\sharp = Z\Sigma^\sharp U^\dagger,$$

where $\Sigma^\sharp$ is the transpose of the matrix obtained from the matrix Σ in the SVD by taking the inverse of each of its nonzero entries and leaving unchanged the zero entries. The Moore-Penrose (MP) pseudo-inverse of $A^\dagger$ is

$$(A^\dagger)^\sharp = (A^\sharp)^\dagger = U(\Sigma^\sharp)^T Z^\dagger = U(\Sigma^\dagger)^\sharp Z^\dagger.$$

Ex. 5.16 *Show that $A^\sharp$ is a generalized inverse for A.*

Some important properties of the MP pseudo-inverse are the following:

- 1. $AA^\sharp A = A$,
- 2. $A^\sharp AA^\sharp = A^\sharp$,
- 3. $(A^\sharp A)^\dagger = A^\sharp A$,
- 4. $(AA^\sharp)^\dagger = AA^\sharp$.

The MP pseudo-inverse of an arbitrary M by N matrix A can be used in much the same way as the inverse of nonsingular matrices to find approximate or exact solutions of systems of equations $Ax = b$. The examples in the following exercises illustrate this point.

Ex. 5.17 *If $M > N$ the system $Ax = b$ probably has no exact solution. Show that whenever $A^\dagger A$ is invertible the pseudo-inverse of A is $A^\sharp = (A^\dagger A)^{-1}A^\dagger$ so that the vector $x = A^\sharp b$ is the least squares approximate solution.*

Ex. 5.18 *If $M < N$ the system $Ax = b$ probably has infinitely many solutions. Show that whenever the matrix $AA^\dagger$ is invertible the pseudo-inverse of A is $A^\sharp = A^\dagger(AA^\dagger)^{-1}$, so that the vector $x = A^\sharp b$ is the exact solution of $Ax = b$ closest to the origin; that is, it is the minimum norm solution.*

In general, the vector $A^\sharp b$ is the vector of smallest norm for which $\|Ax - b\|_2$ is minimized; that is, $A^\sharp b$ is the *minimum-norm least-squares* solution for the system $Ax = b$.

5.6.2 An Example of the MP Pseudo-Inverse

The matrix

$$A = \begin{bmatrix} 4 & 8 & 8 \\ 3 & 6 & 6 \end{bmatrix}$$

has MP pseudo-inverse

$$A^\sharp = \begin{bmatrix} 1/3 & 2/3 & 2/3 \\ 2/3 & -2/3 & 1/3 \\ 2/3 & 1/3 & -2/3 \end{bmatrix} \begin{bmatrix} 1/15 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 4/5 & 3/5 \\ 3/5 & -4/5 \end{bmatrix}.$$

5.6.3 Characterizing the MP Pseudo-Inverse

The MP pseudo-inverse is characterized by the four properties listed above. In other words, an N by M matrix X is the MP pseudo-inverse of A if and only if it satisfies the properties

- 1. $AXA = A$,
- 2. $XAX = X$,
- 3. $(XA)^\dagger = XA$,
- 4. $(AX)^\dagger = AX$.

5.6.4 Calculating the MP Pseudo-Inverse

The properties in the previous subsection that characterize the MP pseudo-inverse suggest algorithms for calculating $X = A^\sharp$ without first calculating the SVD. Let $X = A^\sharp$.

Lemma 5.1 *Let $C = XX^\dagger$. Then $CA^\dagger = X$.*

Proof: We have

$$CA^\dagger = XX^\dagger A^\dagger = X(AX)^\dagger = X(AX) = X.$$

∎

Lemma 5.2 *Let $B = A^\dagger AA^\dagger$. Then $B^\dagger C = A$.*

Proof: We have

$$B^\dagger C = AA^\dagger AXX^\dagger = AA^\dagger(AX)X^\dagger = AA^\dagger(AX)^\dagger X^\dagger = AA^\dagger(X^\dagger A^\dagger)X^\dagger$$
$$= AA^\dagger(XAX)^\dagger = AA^\dagger X^\dagger = A(XA)^\dagger = AXA = A.$$

∎

We know, therefore, that there is at least one Hermitian matrix W, namely $W = C$, having the property that $B^\dagger W = A$. We show now that if we have any Hermitian W with $B^\dagger W = A$, then $WA^\dagger = X = A^\sharp$.

Proposition 5.3 *If $B^\dagger W = A$ and $W^\dagger = W$, then $X = A^\sharp = WA^\dagger$.*

Proof: Let $Y = WA^\dagger$. We show first that $(YA)^\dagger = YA$, or, equivalently, $WA^\dagger A = A^\dagger AW$. From $WB = A^\dagger$ we have

$$A^\dagger(AW) = WB(AW) = WA^\dagger AA^\dagger(AW) = WA^\dagger(B^\dagger W) = WA^\dagger A.$$

Therefore, $(YA)^\dagger = YA$. Next, we show that $(AY)^\dagger = AY$. This is trivial, since we have

$$(AY)^\dagger = (AWA^\dagger)^\dagger = AWA^\dagger = AY.$$

Then we show $YAY = Y$. We have

$$Y = WA^\dagger = W(WB) = W(WA^\dagger A)A^\dagger = W(A^\dagger AW)A^\dagger = YAY.$$

Finally, we show that $AYA = A$. Again, this is easy, since

$$AYA = A(WA^\dagger A) = AA^\dagger AW = B^\dagger W = A.$$

This completes the proof of the proposition. ∎

This proposition suggests that we may be able to calculate the MP pseudo-inverse without first finding the SVD. Suppose that we solve the matrix equations $B^\dagger W = A$ and $W^\dagger = W$. Having found W, we form $Y = WA^\dagger = X$. One approach may be to solve iteratively the combined system $B^\dagger W = A$ and $W = \frac{1}{2}(W + W^\dagger)$. We leave it to the interested reader to investigate the feasibility of this idea.

5.7 Principal-Component Analysis and the SVD

The singular-value decomposition has many uses. One of the most important is as a tool for revealing information hidden in large amounts of data. A good illustration of this is *principal-component analysis* (PCA).

5.7.1 An Example

Suppose, for example, that D is an M by N matrix, that each row of D corresponds to particular applicant to the university, and that each column of D corresponds to a particular measurement of a student's ability or aptitude. One column of D could be SAT mathematics score, another could be IQ, and so on. To permit cross-measurement correlation, the actual scores are not stored, but only the difference between the actual score and the group average; if the average IQ for the group is 110 and John has an IQ of 103, then -7 is entered in the IQ column for John's row. We shall assume that M is greater than N.

The matrix $\frac{1}{M}D^\dagger D$ is the *covariance matrix*, each entry describing how one measurement category is related to a second. We shall focus on the matrix $D^\dagger D$, although proper statistical correlation would require that we normalize to remove the distortions coming from the use of scores that are not all on the same scale. How do we compare twenty points of difference in IQ with one hundred points of difference in SAT score? Once we have calculated $D^\dagger D$, we may find that this N by N matrix is not diagonal, meaning that there is correlation between different measurement categories.

Although the column space of D, denoted $CS(D)$, the span of the columns of D in the space $\mathbb{C}^M$, is probably of dimension N, it may well be the case that the columns of D are nearly spanned by a much smaller set of its members; that is, there is a smaller subset of the columns such that each column of D is nearly equal to a linear combination of the members of this smaller set. That would mean that knowing some of the columns of D, we could predict fairly well what the other columns would be. Statistically speaking, this would say that some scores are highly correlated with others. The goal of principal-component analysis is to find such a smaller set.

5.7.2 Decomposing $D^\dagger D$

The matrix $Q = D^\dagger D$ is Hermitian and non-negative definite; almost certainly, all of its eigenvalues are positive. We list these eigenvalues as follows:

$$\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_N > 0,$$

and assume that λ_{J+k} is nearly zero, for $k = 1, 2, ..., N - J$. With u^j, $j = 1, ..., J$ denoting the orthonormal eigenvectors of $D^\dagger D$ corresponding to the first J eigenvalues, we see that the matrix $D^\dagger D$ is nearly equal to the sum of J dyads:

$$D^\dagger D \approx \sum_{j=1}^{J} \lambda_j u^j (u^j)^\dagger. \tag{5.9}$$

5.7.3 Decomposing D Itself

Let E be the N by J matrix whose J columns are the vectors u^j and R be the J by J diagonal matrix whose entries are $\lambda_j^{-1/2}$, for $j = 1, ..., J$. Let W be the M by J matrix $W = DER$. The matrix D is then approximately equal to the sum of J dyads:

$$D \approx \sum_{j=1}^{J} \sqrt{\lambda_j} w^j (u^j)^\dagger, \tag{5.10}$$

where w^j denotes the jth column of the matrix W. The approximation is with respect to the Frobenius norm. The columns of W lie in $CS(D)$ and each column of D is nearly in the span of the w^j. The w^j are the *principal-component vectors*.

5.7.4 Using the SVD in PCA

In the previous subsection, we obtained a decomposition of the matrix D using the eigenvectors and eigenvalues of the Hermitian matrix $D^\dagger D$. This is not an efficient way to proceed. Instead, we can use the SVD.

Let $A = D^\dagger$. As we saw previously, the singular-value decomposition of A is

$$A = U \Sigma Z^\dagger,$$

so that the SVD of the matrix D is

$$D = Z \Sigma^\dagger U^\dagger = \sum_{j=1}^{N} \sqrt{\lambda_j} z^j (u^j)^\dagger.$$

The first J columns of the matrix Z are the w^j defined above, so the Golub-Kahan SVD algorithm [149] can then be used to obtain the principal-component vectors of the data matrix D.

5.8 The PCA and Factor Analysis

Principal-component analysis has as one of its goals the approximation of a covariance matrix $D^\dagger D$ by nonnegative-definite matrices of lower rank. A related area is *factor analysis*, which attempts to describe an arbitrary N by N Hermitian positive-definite matrix Q as $Q = G^\dagger G + K$, where G is some N by J matrix, for some $J < N$, and K is diagonal. Factor analysis views Q as a covariance matrix, $Q = E(vv^\dagger)$, where v is a random

column vector with mean zero, and attempts to account for the off-diagonal correlated components of Q using the lower-rank matrix $G^\dagger G$. Underlying this is the following model for the random vector v:

$$v = Gx + w,$$

where both x and w are uncorrelated. The entries of the random vector x are the *common factors* that affect each entry of v while those of w are the *special factors*, each associated with a single entry of v. Factor analysis plays an increasingly prominent role in signal and image processing [36] as well as in the social sciences.

In [246] Gil Strang points out that, from a linear algebra standpoint, factor analysis raises some questions. As his example shows, the representation of Q as $Q = G^\dagger G + K$ is not unique. The matrix Q does not uniquely determine the size of the matrix G:

$$Q = \begin{bmatrix} 1 & .74 & .24 & .24 \\ .74 & 1 & .24 & .24 \\ .24 & .24 & 1 & .74 \\ .24 & .24 & .74 & 1 \end{bmatrix} = \begin{bmatrix} .7 & .5 \\ .7 & .5 \\ .7 & -.5 \\ .7 & -.5 \end{bmatrix} \begin{bmatrix} .7 & .7 & .7 & .7 \\ .5 & .5 & -.5 & -.5 \end{bmatrix} + .26I$$

and

$$Q = \begin{bmatrix} .6 & \sqrt{.38} & 0 \\ .6 & \sqrt{.38} & 0 \\ .4 & 0 & \sqrt{.58} \\ .4 & 0 & \sqrt{.58} \end{bmatrix} \begin{bmatrix} .6 & .6 & .4 & .4 \\ \sqrt{.38} & \sqrt{.38} & 0 & 0 \\ 0 & 0 & \sqrt{.58} & \sqrt{.58} \end{bmatrix} + .26I.$$

It is also possible to represent Q with different diagonal components K.

5.9 The MUSIC Method

The "multiple signal identification and classification" (MUSIC) method, originally due to Schmidt [234], is similar to PCA in some respects.

The basic problem now is the following. We have a positive-definite N by N matrix R that we believe has the form

$$R = \sum_{j=1}^{J} \alpha_j e^j (e^j)^\dagger + \sigma^2 I = S + \sigma^2 I, \tag{5.11}$$

where $J < N$ is not known, and the scalars σ and $\alpha_j > 0$, and the column

vectors e^j are not known, but are assumed to be linearly independent. The problem is to determine these unknown scalars and vectors. In applications we usually do have a model for the vectors e^j: it is assumed that each e^j has the form $e^j = e(\theta_j)$, where θ_j is an unknown member of a known family of parameters denoted by θ.

We can say that $R = G^\dagger G + K$, where now $K = \sigma^2 I$, so the MUSIC problem fits into the formulation of factor analysis also. But the MUSIC does more than find a G; it uses the model of parameterized vectors $e(\theta)$ to determine the individual e^j.

The MUSIC method proceeds as follows. First, we calculate the eigenvector/eigenvalue decomposition of R. Let $\lambda_1 \geq ... \geq \lambda_N > 0$ be the ordered eigenvalues, with associated orthonormal eigenvectors u^j. Since $J < N$, we know that the rank of S is J, so that the system $Sx = 0$ has $N - J$ linearly independent solutions. Each of these is an eigenvector of S corresponding to the eigenvalue 0. Therefore, they are also eigenvectors of R corresponding to the eigenvalue $\lambda = \sigma^2$. Since, for $j = 1, 2, ..., J$, $Su^j \neq 0$, for these j we have $\lambda_j > \sigma^2$. So we can tell what J is from the list of eigenvalues of R. Now we find the θ_j. Note that the e^j are in the span of the $u^1, ..., u^J$, but they are not the u^j themselves, generally, since the e^j are probably not mutually orthogonal.

For each $m = 1, ..., N - J$ and each $j = 1, ..., J$, the eigenvector u^{J+m} is orthogonal to e^j. Therefore, the function of θ given by

$$F(\theta) = \sum_{m=1}^{N-J} |(u^{J+m})^\dagger e(\theta)|^2 \tag{5.12}$$

is such that $F(\theta_j) = 0$, for $j = 1, ..., J$. In most situations $F(\theta)$ will have precisely J zeros in the parameter family, so the zeros of $F(\theta)$ will identify the parameter values θ_j. Finding these parameter values then amounts to determining approximately the zeros of $F(\theta)$. Once J and the θ_j have been found, determining the coefficients α_j becomes a linear problem.

5.10 Singular Values of Sparse Matrices

In image reconstruction from projections the M by N matrix A is usually quite large and often ϵ-sparse; that is, most of its elements do not exceed ϵ in absolute value, where ϵ denotes a small positive quantity.

In transmission tomography each column of A corresponds to a single pixel in the digitized image, while each row of A corresponds to a line segment through the object, along which an x-ray beam has traveled. The

entries of a given row of A are nonzero only for those columns whose associated pixel lies on that line segment; clearly, most of the entries of any given row of A will then be zero.

In emission tomography the I by J nonnegative matrix P has entries $P_{ij} \geq 0$; for each detector i and pixel j, P_{ij} is the probability that an emission at the jth pixel will be detected at the ith detector. When a detection is recorded at the ith detector, we want the likely source of the emission to be one of only a small number of pixels. For single photon emission tomography (SPECT), a lead collimator is used to permit detection of only those photons approaching the detector straight on. In positron emission tomography (PET), coincidence detection serves much the same purpose. In both cases the probabilities P_{ij} will be zero (or nearly zero) for most combinations of i and j. Such matrices are called *sparse* (or *almost sparse*).

We discuss now a convenient estimate for the largest singular value of an almost sparse matrix A, which, for notational convenience only, we take to be real. Related estimates of the largest singular value will be presented later, in the chapter "Eigenvalue Bounds in Iteration" .

In [62] it was shown that if A is normalized so that each row has length one, then the spectral radius of A^TA, which is the square of the largest singular value of A itself, does not exceed the maximum number of nonzero elements in any column of A. A similar upper bound on $\rho(A^TA)$ can be obtained for non-normalized, ϵ-sparse A.

Let A be an M by N matrix. For each $n = 1, ..., N$, let $s_n > 0$ be the number of nonzero entries in the nth column of A, and let s be the maximum of the s_n. Let G be the M by N matrix with entries

$$G_{mn} = A_{mn}/(\sum_{l=1}^{N} s_l A_{ml}^2)^{1/2}.$$

Lent has shown that the eigenvalues of the matrix G^TG do not exceed one [195]. This result suggested the following proposition, whose proof was given in [62].

Proposition 5.4 *Let A be an M by N matrix. For each $m = 1, ..., M$ let $\nu_m = \sum_{n=1}^{N} A_{mn}^2 > 0$. For each $n = 1, ..., N$ let $\sigma_n = \sum_{m=1}^{M} e_{mn}\nu_m$, where $e_{mn} = 1$ if $A_{mn} \neq 0$ and $e_{mn} = 0$ otherwise. Let σ denote the maximum of the σ_n. Then the eigenvalues of the matrix A^TA do not exceed σ. If A is normalized so that the Euclidean length of each of its rows is one, then the eigenvalues of A^TA do not exceed s, the maximum number of nonzero elements in any column of A.*

Proof: For simplicity, we consider only the normalized case; the proof for the more general case is similar.

Let $A^TAv = cv$ for some nonzero vector v. We show that $c \leq s$. We have

$AA^TAv = cAv$ and so $w^TAA^Tw = v^TA^TAA^TAv = cv^TA^TAv = cw^Tw$, for $w = Av$. Then, with $e_{mn} = 1$ if $A_{mn} \neq 0$ and $e_{mn} = 0$ otherwise, we have

$$(\sum_{m=1}^{M} A_{mn}w_m)^2 = (\sum_{m=1}^{M} A_{mn}e_{mn}w_m)^2$$

$$\leq (\sum_{m=1}^{M} A_{mn}^2 w_m^2)(\sum_{m=1}^{M} e_{mn}^2) =$$

$$(\sum_{m=1}^{M} A_{mn}^2 w_m^2)s_j \leq (\sum_{m=1}^{M} A_{mn}^2 w_m^2)s.$$

Therefore,

$$w^TAA^Tw = \sum_{n=1}^{N}(\sum_{m=1}^{M} A_{mn}w_m)^2 \leq \sum_{n=1}^{N}(\sum_{m=1}^{M} A_{mn}^2 w_m^2)s,$$

and

$$w^TAA^Tw = c\sum_{m=1}^{M} w_m^2 = c\sum_{m=1}^{M} w_m^2(\sum_{n=1}^{N} A_{mn}^2)$$

$$= c\sum_{m=1}^{M}\sum_{n=1}^{N} w_m^2 A_{mn}^2.$$

The result follows immediately. ∎

If we normalize A so that its rows have length one, then the trace of the matrix AA^T is $\mathrm{tr}(AA^T) = M$, which is also the sum of the eigenvalues of A^TA. Consequently, the maximum eigenvalue of A^TA does not exceed M; this result improves that upper bound considerably, if A is sparse and so $s << M$. A more general theorem along the same lines is Theorem 27.5.

In image reconstruction from projection data that includes scattering we often encounter matrices A most of whose entries are small, if not exactly zero. A slight modification of the proof provides us with a useful upper bound for L, the largest eigenvalue of A^TA, in such cases. Assume that the rows of A have length one. For $\epsilon > 0$ let s be the largest number of entries in any column of A whose magnitudes exceed ϵ. Then we have

$$L \leq s + MN\epsilon^2 + 2\epsilon(MNs)^{1/2}.$$

The proof of this result is similar to that for Proposition 5.4.

5.11 The "Matrix Inversion Theorem"

In this section we bring together several of the conditions equivalent to saying that an N by N matrix A is invertible. Taken together, these conditions are sometimes called the "Matrix Inversion Theorem". The equivalences on the list are roughly in increasing order of difficulty of proof. The reader is invited to supply proofs. We begin with the definition of invertibility.

- 1. According to the definition of invertibility, we say A is invertible if there is a matrix B such that $AB = BA = I$. Then $B = A^{-1}$, the inverse of A.
- 2. A is invertible if and only if there are matrices B and C such that $AB = CA = I$. Then $B = C = A^{-1}$.
- 3. A is invertible if and only if the rank of A is N.
- 4. A is invertible if and only if there is a matrix B with $AB = I$. Then $B = A^{-1}$.
- 5. A is invertible if and only if the columns of A are linearly independent.
- 6. A is invertible if and only if $Ax = 0$ implies $x = 0$.
- 7. A is invertible if and only if A can be transformed by elementary row operations into an upper triangular matrix having no zero entries on its main diagonal.
- 8. A is invertible if and only if the upper triangular matrix $T = U^\dagger AU$ given by Schur's Lemma is invertible, and if and only if there are no zeros on the main diagonal of T.
- 9. A is invertible if and only if its determinant is not zero.
- 10. A is invertible if and only if A has no zero eigenvalues.

5.12 Matrix Diagonalization and Systems of Linear ODE's

We know that the ordinary linear differential equation

$$x'(t) = ax(t)$$

has the solution

$$x(t) = x(0)e^{at}.$$

In this section we use matrix diagonalization to generalize this solution to systems of linear ordinary differential equations.

Consider the system of linear ordinary differential equations

$$x'(t) = 4x(t) - y(t) \tag{5.13}$$
$$y'(t) = 2x(t) + y(t), \tag{5.14}$$

which we write as $z'(t) = Az(t)$, with

$$A = \begin{bmatrix} 4 & -1 \\ 2 & 1 \end{bmatrix},$$

$$z(t) = \begin{bmatrix} x(t) \\ y(t) \end{bmatrix},$$

and

$$z'(t) = \begin{bmatrix} x'(t) \\ y'(t) \end{bmatrix}.$$

We then have

$$\det(A - \lambda I) = (4 - \lambda)(1 - \lambda) + 2 = (\lambda - 2)(\lambda - 3),$$

so the eigenvalues of A are $\lambda = 2$ and $\lambda = 3$.

The vector u given by

$$u = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

solves the system $Au = 2u$ and the vector v given by

$$v = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

solves the system $Av = 3v$. Therefore, u and v are linearly independent eigenvectors of A. With

$$B = \begin{bmatrix} 1 & 1 \\ 2 & 1 \end{bmatrix},$$

$$B^{-1} = \begin{bmatrix} -1 & 1 \\ 2 & -1 \end{bmatrix},$$

and

$$D = \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix},$$

we have $A = BDB^{-1}$ and $B^{-1}AB = D$; this is a diagonalization of A using its eigenvalues and eigenvectors.

Note that not every N by N matrix A will have such a diagonalization; we need N linearly independent eigenvectors of A, which need not exist. They do exist if the eigenvalues of A are all different, as in the example here, and also if the matrix A is Hermitian or normal. The reader should prove that matrix

$$M = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$$

has no such diagonalization.

Continuing with our example, we let $w(t) = B^{-1}z(t)$ so that $w'(t) = Dw(t)$. Because D is diagonal, this new system is uncoupled;

$$w_1'(t) = 2w_1(t),$$

and

$$w_2'(t) = 3w_2(t).$$

The solutions are then

$$w_1(t) = w_1(0)e^{2t},$$

and

$$w_2(t) = w_2(0)e^{3t}.$$

It follows from $z(t) = Bw(t)$ that

$$x(t) = w_1(0)e^{2t} + w_2(0)e^{3t},$$

and

$$y(t) = 2w_1(0)e^{2t} + w_2(0)e^{3t}.$$

We want to express $x(t)$ and $y(t)$ in terms of $x(0)$ and $y(0)$. To do this we use $z(0) = Bw(0)$, which tells us that

$$x(t) = (-x(0) + y(0))e^{2t} + (2x(0) - y(0))e^{3t},$$

and

$$y(t) = (-2x(0) + 2y(0))e^{2t} + (2x(0) - y(0))e^{3t}.$$

We can rewrite this as

$$z(t) = E(t)z(0),$$

where

$$E(t) = \begin{bmatrix} -e^{2t} + 2e^{3t} & e^{2t} - e^{3t} \\ -2e^{2t} + 2e^{3t} & 2e^{2t} - e^{3t} \end{bmatrix}.$$

What is the matrix $E(t)$?

To mimic the solution $x(t) = x(0)e^{at}$ of the problem $x'(t) = ax(t)$, we try

$$z(t) = e^{tA}z(0),$$

with the matrix exponential defined by

$$e^{tA} = \sum_{n=0}^{\infty} \frac{1}{n!} t^n A^n.$$

Since $A = BDB^{-1}$, it follows that $A^n = BD^nB^{-1}$, so that

$$e^{tA} = Be^{tD}B^{-1}.$$

Since D is diagonal, we have

$$e^{tD} = \begin{bmatrix} e^{2t} & 0 \\ 0 & e^{3t} \end{bmatrix}.$$

A simple calculation shows that

$$e^{tA} = B \begin{bmatrix} e^{2t} & 0 \\ 0 & e^{3t} \end{bmatrix} B^{-1} = \begin{bmatrix} -e^{2t} + 2e^{3t} & e^{2t} - e^{3t} \\ -2e^{2t} + 2e^{3t} & 2e^{2t} - e^{3t} \end{bmatrix} = E(t).$$

Therefore, the solution of the original system is

$$z(t) = e^{tA} z(0).$$

5.13 Classical Lie Algebras

Any additive group of square matrices that is closed under the *commutation operation* $[A, B] = AB - BA$ is a matrix Lie (pronounced "Lee") algebra. Here are some examples. Unless otherwise noted, the entries can be real or complex.

- **1.** The collection M_N of all N by N matrices.
- **2.** The collection of matrices in M_N with zero trace.
- **3.** The collection of all real skew-symmetric matrices in M_N.
- **4.** The collection of all A in M_N with $A + A^\dagger = 0$.

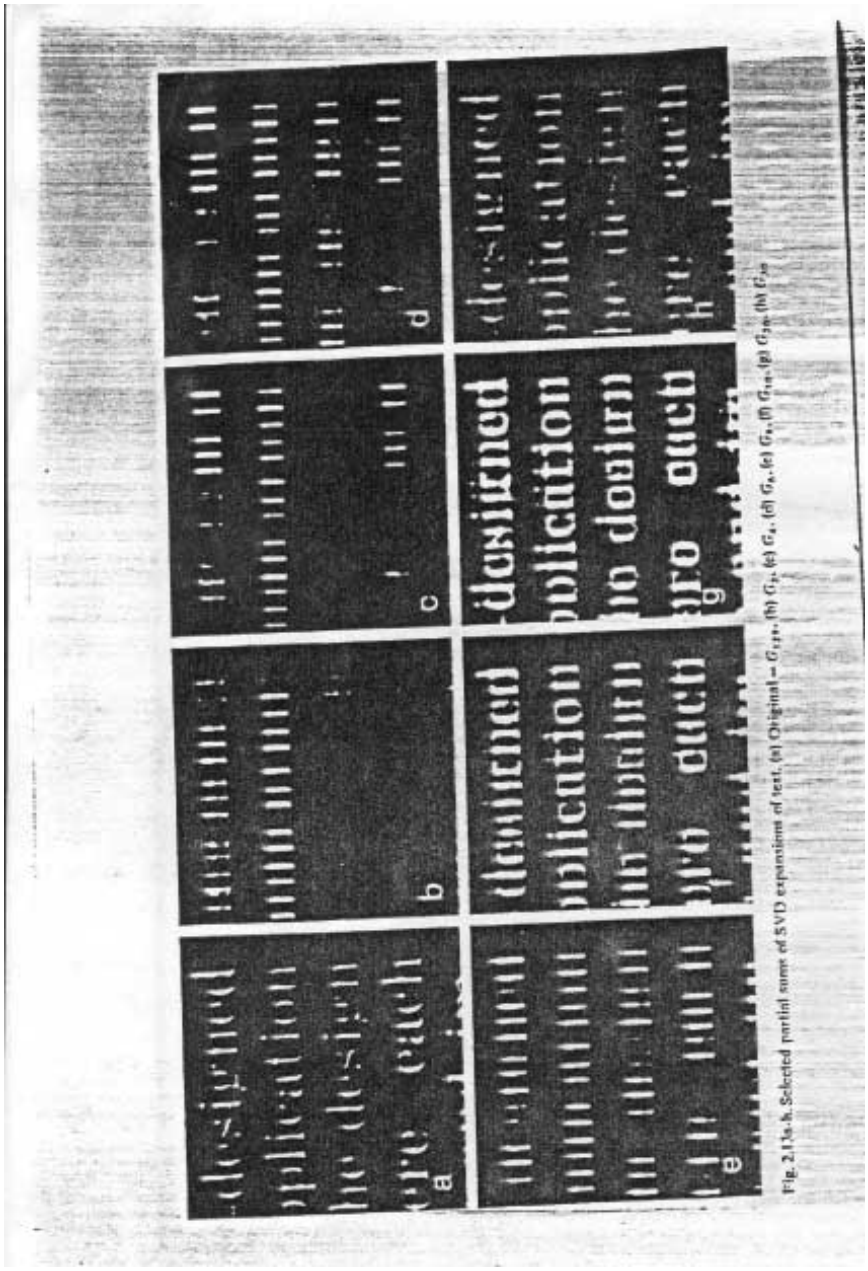

FIGURE 5.1: Compressing text with the SVD.

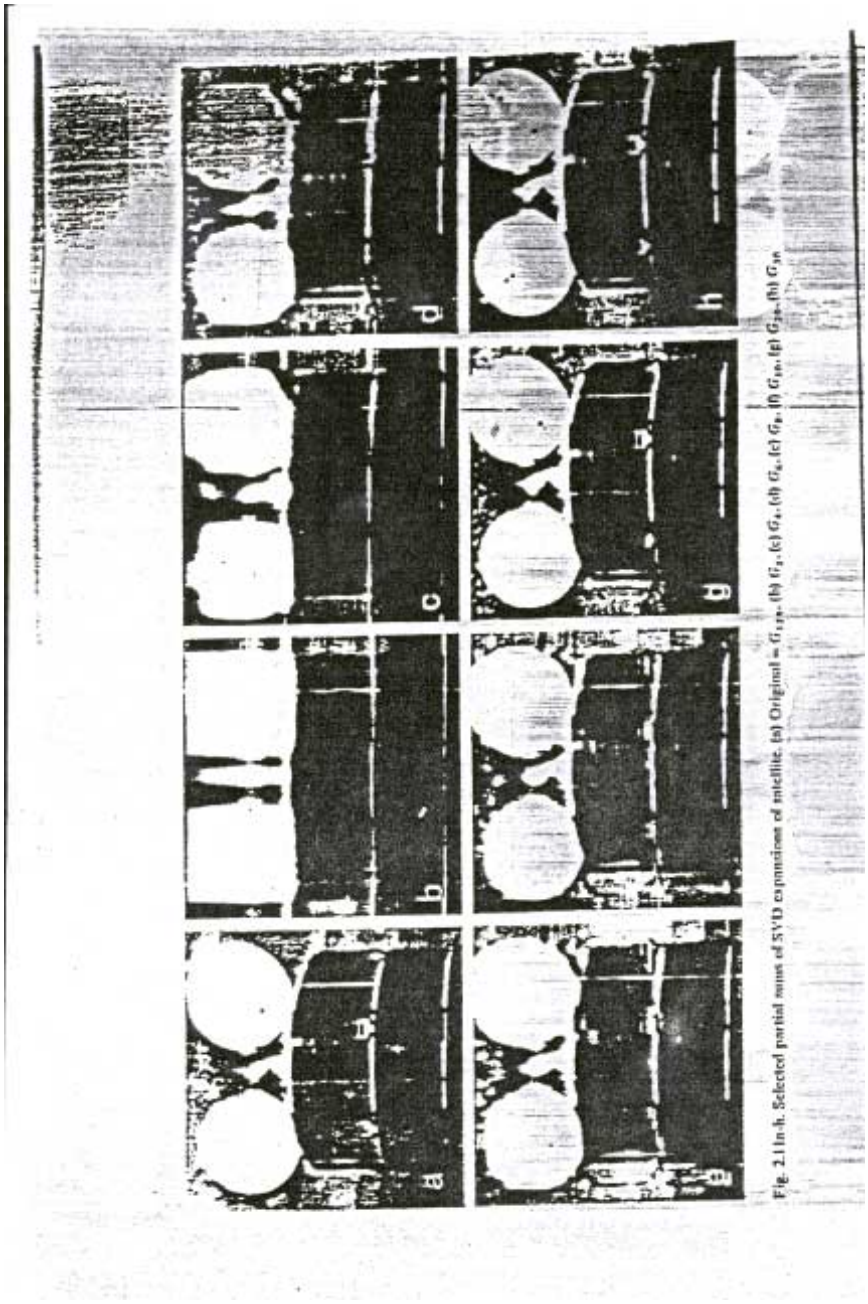

FIGURE 5.2: Compressing an image with the SVD.

Chapter 6

Metric Spaces and Norms

6.1 Chapter Summary

In many applications in which we seek a solution of a linear system of equations $Ax = b$ the entries of the vector b are measurements. If small changes in b result in large changes in the solution x, then we have an unstable situation. In order to measure such changes we need a notion of size of a vector. This leads us to study metrics and norms.

The usual dot product is an inner product on $\mathbb{R}^J$ or $\mathbb{C}^J$ and can be used to define the Euclidean norm $\|x\|_2$ of a vector x, which, in turn, provides a *metric*, or a measure of distance between two vectors, $d(x, y) = \|x - y\|_2$. The notions of metric and norm are actually more general notions, with no necessary connection to the inner product.

6.2 Metric Space Topology

To prepare for our discussion of norms on vectors and matrices we take a quick look at metric space topology.

6.2.1 General Topology

Let S be a non-empty set and $\mathcal{T}$ a non-empty collection of subsets of S. The collection $\mathcal{T}$ is called a *topology* for S if the following conditions hold:

- 1. the empty set and the set S are in $\mathcal{T}$;
- 2. for any finite or infinite sub-collection of members of $\mathcal{T}$, their union is again in $\mathcal{T}$;
- 3. for any positive integer N and sets U_n, $n = 1, 2, ..., N$ in $\mathcal{T}$, their intersection, the set $\cap_{n=1}^{N} U_n$, is in $\mathcal{T}$.

The members of $\mathcal{T}$ are then called the *open sets* for the topology. Notice that we are not given any property that a subset of S may or may not have such that having it would qualify the subset to be called open; a subset of S is open precisely when it is a member of the topology, that is, when it is a member of the collection of subsets called the open subsets. The empty set and S itself are always open, but there need not be any other open subsets. On the other hand, it could be the case that every subset of S is open. It all depends on the collection $\mathcal{T}$ we are given. The *interior* of a subset C of S is the largest open subset of S that is contained within C.

A subset C of S is called a *closed* subset if its complement, the set of all members of S that are not in C, is an open set. The *closure* of a subset C is the smallest closed subset of S that contains C. Once again, we do not describe what it means to be a closed set in terms of some property that C may or may not have, except that its complement is open.

Although the terminology sounds familiar and is borrowed from geometry, these definitions are quite abstract and it is remarkable that a deep theory of topological spaces and continuous functions can be built on such definitions.

6.2.2 Metric Spaces

Metric spaces are the most important and most familiar examples of topological spaces. In contrast to what happens in general topology, now the fundamental notion is that of a *metric* and sets are called open or closed depending on how they behave with respect to the metric. Unlike the general case, now the topology is built up by defining what it means for an individual subset to be open and then including all such subsets in the topology $\mathcal{T}$. We begin with the basic definitions.

Definition 6.1 *Let S be a non-empty set. We say that the function $d : S \times S \to [0, +\infty)$ is a* metric *if the following hold:*

$$d(s,t) \geq 0, \tag{6.1}$$

for all s and t in S;

$$d(s,t) = 0 \tag{6.2}$$

if and only if $s = t$;

$$d(s,t) = d(t,s), \tag{6.3}$$

for all s and t in S; and, for all s, t, and u in S,

$$d(s,t) \leq d(s,u) + d(u,t). \tag{6.4}$$

The pair $\{S, d\}$ is a metric space.

The last inequality is the *Triangle Inequality* for this metric.

6.3 Analysis in Metric Space

Analysis is concerned with issues of convergence and limits.

Definition 6.2 *A sequence* $\{s^k\}$, $k = 1, 2, ...$, *in the metric space* (S, d) *is said to have* limit s^* *if*

$$\lim_{k \to +\infty} d(s^k, s^*) = 0. \tag{6.5}$$

Any sequence with a limit is said to be convergent.

Ex. 6.1 *Show that a sequence can have at most one limit.*

Definition 6.3 *The sequence* $\{s^k\}$ *is said to be a* Cauchy sequence *if, for any* $\epsilon > 0$, *there is positive integer* m, *such that, for any nonnegative integer* n,

$$d(s^m, s^{m+n}) \leq \epsilon. \tag{6.6}$$

Ex. 6.2 *Show that every convergent sequence is a Cauchy sequence.*

Definition 6.4 *The metric space* (S, d) *is said to be* complete *if every Cauchy sequence is a convergent sequence.*

Completeness is part of the axiomatic approach to the definition of the real numbers. From that, it follows that the finite-dimensional spaces $\mathbb{R}^J$and $\mathbb{C}^J$ are complete metric spaces, with respect to the usual Euclidean distance.

Ex. 6.3 *Let* S *be the set of rational numbers, with* $d(s, t) = |s - t|$. *Show that* (S, d) *is a metric space, but not a complete metric space.*

Definition 6.5 *A sequence* $\{s^k\}$ *in* S *is said to be* bounded *if there is a positive constant* $b > 0$ *such that* $d(s^1, s^k) \leq b$, *for all* k.

Ex. 6.4 *Show that any convergent sequence in a metric space is bounded. Find a bounded sequence of real numbers that is not convergent.*

Ex. 6.5 *Show that, if* $\{s^k\}$ *is bounded, then, for any element* c *in the metric space, there is a constant* $r > 0$, *with* $d(c, s^k) \leq r$, *for all* k.

Definition 6.6 *A point* s *in* S *is a* limit point *of a subset* C *of* S *if there are members* c^k *of* C *such that the sequence* $\{c^k\}$ *converges to* s. *Denote by* C^* *the set of all limit points of the set* C.

For any c in C the constant sequence formed by taking $c^k = c$ for each k converges to c. Therefore, every point of C is a limit point of C and $C \subseteq C^*$.

Definition 6.7 *A subset* C *of the metric space is said to be* closed *if every limit point of* C *is in* C; *that is,* $C = C^*$. *The* closure *of a subset* C, *denoted* $cl(C)$, *is the smallest closed set containing* C.

For example, in $\mathbb{R}^J = \mathbb{R}$, the set $C = (0, 1]$ is not closed, because it does not contain the point $s = 0$, which is the limit of the sequence $\{s^k = \frac{1}{k}\}$; the set $C = [0, 1]$ is closed and is the *closure* of the set $(0, 1]$, that is, it is the smallest closed set containing $(0, 1]$.

It is not obvious that there is always a smallest closed set containing C, so it is not clear that the closure of C is well defined. The following proposition gives an explicit description of the closure of C.

Proposition 6.1 *For any subset C of S the closure of C is the set C^*.*

This proposition tells us that we obtain the closure of C by including all its limit points.

Ex. 6.6 *Prove Proposition 6.1. Hint: you need to show that the set C^* is a closed set, which is not immediately obvious. If you think it is obvious, think again.*

Definition 6.8 *For any bounded sequence $\{x^k\}$ in $\mathbb{R}^J$, there is at least one subsequence, often denoted $\{x^{k_n}\}$, that is convergent; the notation implies that the positive integers k_n are ordered, so that $k_1 < k_2 < ....$ The limit of such a subsequence is then said to be a* cluster point *of the original sequence.*

Ex. 6.7 *Show that your bounded, but not convergent, sequence found in Exercise 6.4 has a cluster point.*

Ex. 6.8 *Show that, if x is a cluster point of the sequence $\{x^k\}$, and if $d(x, x^k) \geq d(x, x^{k+1})$, for all k, then x is the limit of the sequence.*

6.4 Motivating Norms

We turn now to metrics that come from norms. Our interest in norms for vectors and matrices stems from their usefulness in analyzing iterative algorithms. Most of the algorithms we shall study involve generating a sequence of vectors $\{x^k\}, k = 0, 1, 2, ...$ in $\mathbb{R}^J$ or $\mathbb{C}^J$, where x^{k+1} comes from x^k according to the formula $x^{k+1} = T(x^k)$, where T is a (possibly nonlinear) operator on the space of vectors. When we investigate iterative algorithms, we will want to know if the sequence $\{x^k\}$ generated by the algorithm converges. As a first step, we will usually ask if the sequence is bounded? If it is bounded, then it will have at least one cluster point. We then try to discover if that cluster point is really the limit of the sequence.

It would help if we know that the vector $T(x) - T(y)$ is smaller, in some sense, than the vector $x - y$.

Affine operators T have the form $T(x) = Bx + d$, where B is a matrix and d is a fixed vector. Such affine operators arise, for example, in the Landweber algorithm for solving $Ax = b$; the iterative step is

$$x^{k+1} = x^k + \gamma A^\dagger(b - (Ax^k)),$$

which we can write as

$$x^{k+1} = (I - \gamma A^\dagger A)x^k + \gamma A^\dagger b.$$

Then $x^{k+1} = T(x^k)$, where T is the affine operator

$$T(x) = (I - \gamma A^\dagger A)x + \gamma A^\dagger b.$$

For affine operators $T(x) - T(y) = Bx - By = B(x - y)$, so we are interested in the size of Bz, relative to the size of z, for all vectors z. Vector and matrix norms will help us here.

6.5 Norms

The metric spaces that interest us most are vector spaces V for which the metric comes from a norm, which is a measure of the length of a vector.

Definition 6.9 *We say that $\|\cdot\|$ is a* norm *on V if*

$$\|x\| \geq 0, \tag{6.7}$$

for all x,

$$\|x\| = 0 \tag{6.8}$$

if and only if $x = 0$,

$$\|\gamma x\| = |\gamma|\,\|x\|, \tag{6.9}$$

for all x and scalars γ, and

$$\|x + y\| \leq \|x\| + \|y\|, \tag{6.10}$$

for all vectors x and y.

Lemma 6.1 *The function $d(x, y) = \|x - y\|$ defines a metric on V.*

It can be shown that $\mathbb{R}^J$ and $\mathbb{C}^J$ are complete for any metric arising from a norm.

6.5.1 Some Common Norms on $\mathbb{C}^J$

We consider now the most common norms on the space $\mathbb{C}^J$. These notions apply equally to $\mathbb{R}^J$.

6.5.1.1 The 1-norm

The 1-norm on $\mathbb{C}^J$ is defined by

$$\|x\|_1 = \sum_{j=1}^{J} |x_j|. \tag{6.11}$$

6.5.1.2 The ∞-norm

The ∞-norm on $\mathbb{C}^J$ is defined by

$$\|x\|_\infty = \max\{|x_j| \, | j = 1, ..., J\}. \tag{6.12}$$

6.5.1.3 The p-norm

For any $p \geq 1$, the p-norm is defined by

$$\|x\|_p = \Big(\sum_{j=1}^{J} |x_j|^p\Big)^{1/p}. \tag{6.13}$$

6.5.1.4 The 2-norm

The 2-norm, also called the Euclidean norm, is the most commonly used norm on $\mathbb{C}^J$. It is the p-norm for $p = 2$ and is the one that comes from the inner product:

$$\|x\|_2 = \sqrt{\sum_{j=1}^{J} |x_j|^2} = \sqrt{\langle x, x\rangle} = \sqrt{x^\dagger x}. \tag{6.14}$$

6.5.1.5 Weighted 2-norms

Let A be an invertible matrix and $Q = A^\dagger A$. Define

$$\|x\|_Q = \|Ax\|_2 = \sqrt{x^\dagger Q x}, \tag{6.15}$$

for all vectors x. This is the *Q-weighted* 2-norm of x. If Q is the diagonal matrix with diagonal entries $Q_{jj} > 0$, then

$$\|x\|_Q = \sqrt{\sum_{j=1}^{J} Q_{jj}|x_j|^2}. \tag{6.16}$$

Ex. 6.9 *Show that the* 1*-norm is a norm.*

Ex. 6.10 *Show that the* ∞*-norm is a norm.*

Ex. 6.11 *Show that the* 2*-norm is a norm. Hint: for the triangle inequality, use the Cauchy Inequality.*

Ex. 6.12 *Show that the* Q*-weighted* 2*-norm is a norm.*

6.6 The Hölder and Minkowski Inequalities

To show that the p-norm is a norm we need Minkowski's Inequality, which follows from Hölder's Inequality.

Let $c = (c_1, ..., c_N)$ and $d = (d_1, ..., d_N)$ be vectors with complex entries and let p and q be positive real numbers such that

$$\frac{1}{p} + \frac{1}{q} = 1.$$

The p-norm of c is defined to be

$$\|c\|_p = \Big(\sum_{n=1}^{N} |c_n|^p \Big)^{1/p},$$

with the q-norm of d, denoted $\|d\|_q$, defined similarly.

6.6.1 Hölder's Inequality

Hölder's Inequality is the following:

$$\sum_{n=1}^{N} |c_n d_n| \leq \|c\|_p \|d\|_q,$$

with equality if and only if

$$\Big(\frac{|c_n|}{\|c\|_p} \Big)^p = \Big(\frac{|d_n|}{\|d\|_q} \Big)^q,$$

for each n.

Hölder's Inequality follows from the GAGM Inequality. To see this, we fix n and apply Inequality (30.2), with

$$x_1 = \Big(\frac{|c_n|}{\|c\|_p}\Big)^p,$$

$$a_1 = \frac{1}{p},$$

$$x_2 = \Big(\frac{|d_n|}{\|d\|_q}\Big)^q,$$

and

$$a_2 = \frac{1}{q}.$$

From (30.2) we then have

$$\Big(\frac{|c_n|}{\|c\|_p}\Big)\Big(\frac{|d_n|}{\|d\|_q}\Big) \leq \frac{1}{p}\Big(\frac{|c_n|}{\|c\|_p}\Big)^p + \frac{1}{q}\Big(\frac{|d_n|}{\|d\|_q}\Big)^q.$$

Now sum both sides over the index n.

It will be helpful later to note here that

$$\sum_{n=1}^{N} \overline{c_n} d_n = \sum_{n=1}^{N} |c_n||d_n|$$

if each $\overline{c_n}d_n$ is non-negative, which means that the complex numbers c_n and d_n have the same phase angles.

6.6.2 Minkowski's Inequality

Minkowski's Inequality, which is a consequence of Hölder's Inequality, states that

$$\|c + d\|_p \leq \|c\|_p + \|d\|_p\,;$$

it is the triangle inequality for the metric induced by the p-norm.

To prove Minkowski's Inequality, we write

$$\sum_{n=1}^{N} |c_n + d_n|^p \leq \sum_{n=1}^{N} |c_n|(|c_n + d_n|)^{p-1} + \sum_{n=1}^{N} |d_n|(|c_n + d_n|)^{p-1}.$$

Then we apply Hölder's Inequality to both of the sums on the right side of the equation.

For the choices $p = q = 2$, Hölder's Inequality becomes the famous Cauchy Inequality.

Ex. 6.13 *Show that the p-norm is a norm.*

6.7 Matrix Norms

Any matrix can be turned into a vector by vectorization. Therefore, we can define a norm for any matrix by simply vectorizing the matrix and taking a norm of the resulting vector; the 2-norm of the vectorized matrix is the *Frobenius norm* of the matrix itself. Such norms for matrices may not be compatible with the role of a matrix as representing a linear transformation. For that reason, we consider norms on matrices that are induced by the norms of the vectors on which the matrices operate.

Definition 6.10 *Let A be an M by N complex matrix. A norm on A, denoted $\|A\|$, is said to be* compatible *with given norms on $\mathbb{C}^N$ and $\mathbb{C}^M$ if $\|Ax\| \leq \|A\|\|x\|$, for every x in $\mathbb{C}^N$.*

6.7.1 Induced Matrix Norms

One way to obtain a compatible norm for matrices is through the use of an induced matrix norm.

Definition 6.11 *Let $\|x\|$ be any norm on $\mathbb{C}^J$, not necessarily the Euclidean norm, $\|b\|$ any norm on $\mathbb{C}^I$, and A a rectangular I by J matrix. The* induced matrix norm *of A, simply denoted $\|A\|$, derived from these two vector norms, is the smallest positive constant c such that*

$$\|Ax\| \leq c\|x\|, \tag{6.17}$$

for all x in $\mathbb{C}^J$. This induced norm can be written as

$$\|A\| = \max_{x \neq 0}\{\|Ax\|/\|x\|\}. \tag{6.18}$$

When A is square we always assume that it is the same norm being used on x and Ax.

We study induced matrix norms in order to measure the distance from Ax to Az, $\|Ax - Az\|$, relative to $\|x - z\|$, the distance from x to z:

$$\|Ax - Az\| \leq \|A\|\,\|x - z\|, \tag{6.19}$$

for all vectors x and z and $\|A\|$ is the smallest number for which this statement is valid.

Ex. 6.14 *Show that $\rho(S) \leq \|S\|$ for any square matrix S.*

Ex. 6.15 *Let the matrices A be M by N, and B be N by K. Show that, for any norms on the spaces $\mathbb{R}^M$, $\mathbb{R}^N$ and $\mathbb{R}^K$, we have the inequality*

$$\|AB\| \leq \|A\| \, \|B\|,$$

for the induced matrix norms.

Using the next two lemmas, we can show that there are induced matrix norms for S that are as close to $\rho(S)$ as we wish.

Lemma 6.2 *Let M be an invertible matrix and $\|x\|$ any vector norm. Define*

$$\|x\|_M = \|Mx\|. \tag{6.20}$$

Then, for any square matrix S, the matrix norm

$$\|S\|_M = \max_{x \neq 0} \{\|Sx\|_M / \|x\|_M\} \tag{6.21}$$

is

$$\|S\|_M = \|MSM^{-1}\|. \tag{6.22}$$

In [7] this result is used to prove the following lemma:

Lemma 6.3 *Let S be any square matrix and let $\epsilon > 0$ be given. Then there is an invertible matrix M such that*

$$\|S\|_M \leq \rho(S) + \epsilon. \tag{6.23}$$

Later, we shall show that if an N by N matrix S is diagonalizable, that is, if there is a basis for $\mathbb{C}^N$ consisting of eigenvectors of S, then there is an invertible matrix M such that $\|S\|_M = \rho(S)$.

Ex. 6.16 *Show that, if $\rho(S) < 1$, then there is a vector norm on $\mathbb{C}^J$ for which the induced matrix norm of S is less than one.*

Ex. 6.17 *Show that $\rho(S) < 1$ if and only if $\lim_{k \to \infty} S^k = 0$.*

Definition 6.12 *Let A be an arbitrary matrix. Denote by $|A|$ the matrix whose entries are the absolute values of those of A, that is, $|A|_{mn} = |A_{mn}|$.*

Proposition 6.2 *Let A and B be N by N real matrices. If $|A|_{mn} \leq B_{mn}$ for all m and n, then $\rho(A) \leq \rho(B)$.*

Proof: Let $\sigma = \rho(B)$ and $\epsilon > 0$ be arbitrary. Let $B_1 = (\sigma + \epsilon)^{-1} B$ and $A_1 = (\sigma + \epsilon)^{-1} A$. Then $\rho(B_1) < 1$, so that $B_1^k \to 0$, as $k \to \infty$. Therefore, $A_1^k \to 0$ also. From Exercise 6.17 we can conclude that $\rho(A_1) < 1$. Therefore, $\rho(A) < \sigma + \epsilon$. Since ϵ is arbitrary, it follows that $\rho(A) \leq \sigma = \rho(B)$. ∎

Corollary 6.1 *For any square matrix A we have $\rho(A) \leq \rho(|A|)$.*

6.7.2 Some Examples of Induced Matrix Norms

If we choose the two vector norms carefully, then we can get an explicit description of $\|A\|$, but, in general, we cannot.

For example, let $\|x\| = \|x\|_1$ and $\|Ax\| = \|Ax\|_1$ be the 1-norms of the vectors x and Ax, where

$$\|x\|_1 = \sum_{j=1}^{J} |x_j|. \tag{6.24}$$

Lemma 6.4 *The* 1*-norm of A, induced by the* 1*-norms of vectors in* $\mathbb{C}^J$ *and* $\mathbb{C}^I$*, is*

$$\|A\|_1 = \max \{\sum_{i=1}^{I} |A_{ij}|, j = 1, 2, ..., J\}. \tag{6.25}$$

Proof: Use basic properties of the absolute value to show that

$$\|Ax\|_1 \leq \sum_{j=1}^{J} (\sum_{i=1}^{I} |A_{ij}|)|x_j|. \tag{6.26}$$

Then let $j = m$ be the index for which the maximum column sum is reached and select $x_j = 0$, for $j \neq m$, and $x_m = 1$. ∎

The *infinity norm* of the vector x is

$$\|x\|_\infty = \max \{|x_j|, j = 1, 2, ..., J\}. \tag{6.27}$$

Lemma 6.5 *The infinity norm of the matrix A, induced by the infinity norms of vectors in* $\mathbb{R}^J$ *and* $\mathbb{C}^I$*, is*

$$\|A\|_\infty = \max \{\sum_{j=1}^{J} |A_{ij}|, i = 1, 2, ..., I\}. \tag{6.28}$$

The proof is similar to that of the previous lemma.

From these two lemmas we learn that

$$\|A^\dagger\|_1 = \|A\|_\infty,$$

and

$$\|A^\dagger\|_\infty = \|A\|_1.$$

6.7.3 The Two-Norm of a Matrix

We shall be particularly interested in the two-norm (or 2-norm) of a matrix A, denoted by $\|A\|_2$, which is the induced matrix norm derived from the Euclidean vector norms.

From the definition of the two-norm of A, we know that

$$\|A\|_2 = \max\{\|Ax\|_2/\|x\|_2\}, \tag{6.29}$$

with the maximum over all nonzero vectors x. Since

$$\|Ax\|_2^2 = x^\dagger A^\dagger A x, \tag{6.30}$$

we have

$$\|A\|_2 = \sqrt{\max\{\frac{x^\dagger A^\dagger A x}{x^\dagger x}\}}, \tag{6.31}$$

over all nonzero vectors x.

Proposition 6.3 *The two-norm of a matrix A is*

$$\|A\|_2 = \sqrt{\rho(A^\dagger A)}; \tag{6.32}$$

that is, the term inside the square-root in Equation (6.31) is the largest eigenvalue of the matrix $A^\dagger A$.

Proof: Let

$$\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_J \geq 0 \tag{6.33}$$

be the eigenvalues of $A^\dagger A$, and let $\{u^j,\, j = 1, ..., J\}$ be the associated mutually orthogonal eigenvectors of $A^\dagger A$ with $\|u^j\|_2 = 1$. Then, for any x, we have

$$x = \sum_{j=1}^{J} [(u^j)^\dagger x] u^j, \tag{6.34}$$

while

$$A^\dagger A x = \sum_{j=1}^{J} [(u^j)^\dagger x] A^\dagger A u^j = \sum_{j=1}^{J} \lambda_j [(u^j)^\dagger x] u^j. \tag{6.35}$$

It follows that

$$\|x\|_2^2 = x^\dagger x = \sum_{j=1}^{J} |(u^j)^\dagger x|^2, \tag{6.36}$$

and

$$\|Ax\|_2^2 = x^\dagger A^\dagger Ax = \sum_{j=1}^{J} \lambda_j |(u^j)^\dagger x|^2. \tag{6.37}$$

Maximizing $\|Ax\|_2^2/\|x\|_2^2$ over $x \neq 0$ is equivalent to maximizing $\|Ax\|_2^2$, subject to $\|x\|_2^2 = 1$. The right side of Equation (6.37) is then a convex combination of the λ_j, which will have its maximum when only the coefficient of λ_1 is non-zero. ∎

Ex. 6.18 *Show that $\|A\|_2 = \|A^\dagger\|_2$ for any matrix A. Hints: use Exercise 5.12 and Proposition 6.3.*

Note that it can be shown ([7], p. 164) that for any square matrix S and any matrix norm we have

$$\rho(S) = \lim_{n\to\infty} (\|S^n\|)^{1/n}.$$

6.7.4 The Two-norm of an Hermitian Matrix

Let H be an Hermitian matrix. We then have the following result:

Proposition 6.4 *The two-norm of H is $\|H\|_2 = \rho(H)$.*

Ex. 6.19 *Prove Proposition 6.4. Hint: use $H^\dagger H = H^2$ and Exercise 5.8.*

Using Proposition 6.4, we can prove the following theorem.

Theorem 6.1 *For any matrix A we have the inequality*

$$\|A\|_2^2 \leq \|A\|_1 \|A\|_\infty. \tag{6.38}$$

Proof: Let $H = A^\dagger A$. We know that $\|A\|_2^2 = \|H\|_2$ and that

$$\|H\|_2 \leq \|H\|_1 = \|A^\dagger A\|_1 \leq \|A^\dagger\|_1 \|A\|_1 = \|A\|_\infty \|A\|_1.$$

∎

The inequality (6.38) also follows, as a particular case, from the more general Theorem 27.5 concerning upper bounds for the singular values of a matrix A.

Ex. 6.20 *Show that if the rows of the matrix A are rescaled so that, for each i, we have $\sum_{j=1}^{J} |A_{ij}| \leq 1$, then no eigenvalue of $A^\dagger A$ is larger than the maximum number of non-zero entries in any column of A. In Corollary 27.2 we shall see that the same conclusion holds if the rows of A are rescaled to have Euclidean length not greater than one.*

If S is not Hermitian, then the two-norm of S cannot be calculated directly from the eigenvalues of S. Take, for example, the square, non-Hermitian matrix

$$S = \begin{bmatrix} i & 2 \\ 0 & i \end{bmatrix}, \tag{6.39}$$

having eigenvalues $\lambda = i$ and $\lambda = i$. The eigenvalues of the Hermitian matrix

$$S^\dagger S = \begin{bmatrix} 1 & -2i \\ 2i & 5 \end{bmatrix} \tag{6.40}$$

are $\lambda = 3 + 2\sqrt{2}$ and $\lambda = 3 - 2\sqrt{2}$. Therefore, the two-norm of S is

$$\|S\|_2 = \sqrt{3 + 2\sqrt{2}}. \tag{6.41}$$

6.7.5 The p-norm of a Matrix

The p-norm of an I by J complex matrix A is the norm induced by the p-norms on the vectors in $\mathbb{C}^I$ and $\mathbb{C}^J$; we can say that $\|A\|_p$ is the maximum of $\|Ax\|_p$, over all x with $\|x\|_p = 1$.

Previously, we were able to use the explicit descriptions of $\|A\|_1$ and $\|A\|_\infty$ to show that $\|A^\dagger\|_1 = \|A\|_\infty$. A similar result holds for the p-norm.

Theorem 6.2 *Let $\frac{1}{p} + \frac{1}{q} = 1$. Then*

$$\|A^\dagger\|_p = \|A\|_q.$$

Proof: We select a vector x with $\|x\|_p = 1$. We then construct the vector v with

$$|v_i|^q = |(Ax)_i|^p / \|Ax\|_p^p,$$

and such that v_i and $(Ax)_i$ have the same phase angles. Then $\|v\|_q = 1$. It follows that

$$\sum_{i=1}^{I} \overline{(Ax)_i} v_i = \|Ax\|_p.$$

We also have

$$\sum_{i=1}^{I} \overline{(Ax)_i} v_i = \sum_{j=1}^{J} \overline{x_j} (A^\dagger v)_j,$$

so that

$$\|Ax\|_p = \sum_{j=1}^{J} \overline{x_j}(A^\dagger v)_j \leq \|x\|_p \|A^\dagger v\|_q.$$

It then follows that the maximum of $\|Ax\|_p$, over all x with $\|x\|_p = 1$, is not greater than the maximum of $\|A^\dagger v\|_q$, over all v with $\|v\|_q = 1$. Since this is true for all A, the theorem follows. ∎

We can use Theorem 6.2 to prove *Young's Inequality*.

Theorem 6.3 (Young's Inequality) *For any complex matrix A we have*

$$\|A\|_2^2 \leq \|A\|_p \|A\|_q. \tag{6.42}$$

Proof: We know that $\rho(S) \leq \|S\|$, for all square matrices S and all induced matrix norms. Also, for $S = H$ Hermitian, we have $\rho(H) = \|H\|_2$, from which we conclude that $\|H\|_2 \leq \|H\|$, for all induced matrix norms. Now we let $H = A^\dagger A$.

From $\|A\|_2^2 = \|H\|_2$, we have

$$\|A\|_2^2 = \sqrt{\|H\|_2^2} = \sqrt{\|H\|_2 \|H\|_2} \leq \sqrt{\|H\|_p \|H\|_q}.$$

Since

$$\|H\|_p = \|A^\dagger A\|_p \leq \|A^\dagger\|_p \|A\|_p = \|A\|_q \|A\|_p,$$

it follows that

$$\|A\|_2^2 \leq \|A\|_p \|A\|_q.$$

∎

6.7.6 Diagonalizable Matrices

Definition 6.13 *A J by J matrix S is* diagonalizable *if $\mathbb{C}^J$ has a basis of eigenvectors of S.*

As the following lemma tells us, most square matrices are diagonalizable.

Lemma 6.6 *A square matrix S is diagonalizable if all its eigenvalues are distinct.*

Proof: We need to show that the eigenvectors associated with different eigenvalues are linearly independent. Let S be J by J. Let λ_j be the eigenvalues of S, $Su^j = \lambda_j u^j$, and $u^j \neq 0$, for $j = 1, ..., J$. Let u^m be the first eigenvector that is in the span of $\{u_j | j = 1, ..., m-1\}$. Then

$$u^m = a_1 u^1 + ... + a_{m-1} u^{m-1}, \tag{6.43}$$

for some constants a_j that are not all zero. Multiply both sides by λ_m to get

$$\lambda_m u^m = a_1 \lambda_m u^1 + ... + a_{m-1} \lambda_m u^{m-1}. \tag{6.44}$$

From

$$\lambda_m u^m = A u^m = a_1 \lambda_1 u^1 + ... + a_{m-1} \lambda_{m-1} u^{m-1}, \tag{6.45}$$

it follows that

$$a_1(\lambda_m - \lambda_1)u^1 + ... + a_{m-1}(\lambda_m - \lambda_{m-1})u^{m-1} = 0, \tag{6.46}$$

from which we can conclude that some u^n in $\{u^1, ..., u^{m-1}\}$ is in the span of the others. This is a contradiction. ∎

When S is diagonalizable, we let U be a square matrix whose columns are J linearly independent eigenvectors of S and L the diagonal matrix having the eigenvalues of S along its main diagonal; then we have $SU = UL$, or $U^{-1}SU = L$.

Ex. 6.21 *Let $M = U^{-1}$ and define $\|x\|_M = \|Mx\|_2$, the Euclidean norm of Mx. Show that the induced matrix norm of S is $\|S\|_M = \rho(S)$.*

We see from this exercise that, for any diagonalizable matrix S, in particular, for any Hermitian matrix, there is a vector norm such that the induced matrix norm of S is $\rho(S)$.

In the Hermitian case $S = H$, we know that we can select the eigenvector columns of U to be mutually orthogonal and scaled to have length one, so that $U^{-1} = U^\dagger$ and $\|Mx\|_2 = \|U^\dagger x\|_2 = \|x\|_2$, so that the required vector norm is just the Euclidean norm, and $\|H\|_M$ is just $\|H\|_2$, which we know to be $\rho(H)$.

Ex. 6.22 *The* Cayley-Hamilton Theorem *asserts that if S is any square matrix and $P(\lambda)$ its characteristic polynomial, then $P(S) = 0$. Prove this for the case of diagonalizable S.*

6.8 Estimating Eigenvalues

Calculating the eigenvalues of a square matrix amounts to solving for the roots of a polynomial. In general, this requires an iterative procedure, since there are no algebraic formulas for finding the roots of arbitrary polynomials. In this section we give two simply methods for obtaining somewhat crude estimates of the eigenvalues. Later, we shall present better estimation methods.

6.8.1 Using the Trace

The trace of a square matrix S, written trace(S) or tr(S), is the sum of the entries on the main diagonal of S. If S is diagonalizable, then we can write $S = ULU^{-1}$, where L is the diagonal matrix whose diagonal entries are the eigenvalues of S. For any square matrices A, B, and C we have

$$\text{tr}(ABC) = \text{tr}(CAB) = \text{tr}(BCA),$$

but these are not necessarily equal to tr(BAC). Therefore,

$$\text{tr}(S) = \text{tr}(ULU^{-1}) = \text{tr}(U^{-1}UL) = \text{tr}(L),$$

so that the trace of S is the sum of its eigenvalues. The same result holds for non-diagonalizable matrices, but the proof is a bit harder; try to prove this using Schur's Lemma 5.1.

6.8.2 Gerschgorin's Theorem

Gerschgorin's theorem gives us a way to estimate the eigenvalues of an arbitrary square matrix S.

Theorem 6.4 *Let S be J by J. For $j = 1, ..., J$, let C_j be the circle in the complex plane with center S_{jj} and radius $r_j = \sum_{m \neq j} |S_{jm}|$. Then every eigenvalue of S lies within one of the C_j.*

Proof: Let λ be an eigenvalue of S, with associated eigenvector u. Let u_j be the entry of the vector u having the largest absolute value. From $Su = \lambda u$, we have

$$(\lambda - S_{jj})u_j = \sum_{m \neq j} S_{jm}u_m, \tag{6.47}$$

so that

$$|\lambda - S_{jj}| \leq \sum_{m \neq j} |S_{jm}||u_m|/|u_j| \leq r_j. \tag{6.48}$$

This completes the proof. ∎

6.8.3 Strictly Diagonally Dominant Matrices

Definition 6.14 *A square I by I matrix S is said to be* strictly diagonally dominant *if, for each $i = 1, ..., I$,*

$$|S_{ii}| > r_i = \sum_{m \neq i} |S_{im}|. \tag{6.49}$$

When the matrix S is strictly diagonally dominant, all the eigenvalues of S lie within the union of the spheres with centers S_{ii} and radii S_{ii}. With D the diagonal component of S, the matrix $D^{-1}S$ then has all its eigenvalues within the circle of radius one, centered at $(1, 0)$. Then $\rho(I - D^{-1}S) < 1$. This result is used when we discuss the Jacobi splitting method [66].

6.9 Conditioning

6.9.1 Condition Number of a Square Matrix

Let S be a square, invertible matrix and z the solution to $Sz = h$. We are concerned with the extent to which the solution changes as the right side, h, changes. Denote by δ_h a small perturbation of h, and by δ_z the solution of $S\delta_z = \delta_h$. Then $S(z + \delta_z) = h + \delta_h$. Applying the compatibility condition $\|Ax\| \leq \|A\|\|x\|$, we get

$$\|\delta_z\| \leq \|S^{-1}\|\|\delta_h\|, \tag{6.50}$$

and

$$\|z\| \geq \|h\|/\|S\|. \tag{6.51}$$

Therefore

$$\frac{\|\delta_z\|}{\|z\|} \leq \|S\|\,\|S^{-1}\|\frac{\|\delta_h\|}{\|h\|}. \tag{6.52}$$

Definition 6.15 *The quantity $c = \|S\|\|S^{-1}\|$ is the* condition number *of S, with respect to the given matrix norm.*

Note that $c \geq 1$: for any non-zero z, we have

$$1 = \|I\| = \|SS^{-1}\| \leq \|S\|\|S^{-1}\|. \tag{6.53}$$

Ex. 6.23 *Show that when Q is Hermitian and positive-definite, the condition number of Q, with respect to the matrix norm induced by the Euclidean vector norm, is*

$$c = \lambda_{max}(Q)/\lambda_{min}(Q), \tag{6.54}$$

the ratio of the largest to the smallest eigenvalues of Q.

Chapter 7

Under-Determined Systems of Linear Equations

7.1 Chapter Summary

When a system of M linear equations in N unknowns, denoted $Ax = b$, has multiple solutions, we say that the system is *under-determined*. Then it has infinitely many solutions; if $Ax = b$ and $Az = b$ and $x \neq z$, then $x + \alpha(z - x)$ is also a solution, for any scalar α. In such cases, we usually select one solution out of the infinitely many possibilities by requiring that the solution also satisfy some additional constraints. For example, we can select that solution x for which $\|x\|_2$ is minimized, which we denote by $\hat{x}$. This *minimum two-norm* solution is given by

$$\hat{x} = A^\dagger (AA^\dagger)^{-1} b,$$

provided that the matrix $AA^\dagger$ has an inverse. In this chapter we survey several of the constraints that are commonly used and the algorithms that are employed to calculate these constrained solutions.

7.2 Minimum Two-Norm Solutions

When the system $Ax = b$ is under-determined, it is reasonable to ask for that solution $x = \hat{x}$ having the smallest two-norm

$$\|x\|_2 = \sqrt{\sum_{n=1}^{N} |x_n|^2}.$$

As we showed previously, the *minimum two-norm* solution of $Ax = b$ is a vector of the form $\hat{x} = A^\dagger z$. Then $A\hat{x} = b$ becomes $AA^\dagger z = b$. Typically, $(AA^\dagger)^{-1}$ will exist, and we get $z = (AA^\dagger)^{-1}b$, from which it follows that the minimum two-norm solution is $\hat{x} = A^\dagger(AA^\dagger)^{-1}b$. When M and N are not too large, forming the matrix $AA^\dagger$ and solving for z is not prohibitively expensive or time-consuming.

When M and N are large, we turn to iterative algorithms to find the minimum two-norm solution. Both the ART and the Landweber algorithm converge to that solution closest to the starting vector x^0, in the two-norm sense. Therefore, when we begin with $x^0 = 0$, these algorithms give us the minimum two-norm solution.

If C is a closed convex set in $\mathbb{R}^N$, the *projected Landweber algorithm* converges to that solution x in C closest to x^0, in the two-norm sense. Again, if we take $x^0 = 0$, the projected Landweber algorithm converges to that solution x in C having the smallest two-norm.

7.3 Minimum Weighted Two-Norm Solutions

The *minimum weighted two-norm solution* is the $x = \tilde{x}$ satisfying $Ax = b$ for which the weighted two-norm

$$\|x\|_w = \sqrt{\sum_{n=1}^{N} |x_n|^2 w_n}$$

is minimized. This solution can be found easily by changing variables, letting $u_n = x_n\sqrt{w_n}$, to convert the problem into a minimum two-norm

problem, and then applying any of the methods discussed in the previous chapter. The minimum weighted two-norm approach is a discrete version of a method, called the PDFT, for estimating a function from values of its Fourier transform [43].

Figure 2.2 illustrates the potential advantages to be obtained through the use of weights. In that example, we have a prior estimate of the magnitudes of the x_n, which we called $p_n > 0$. Then we chose for the weights $w_n = p_n^{-1}$.

7.4 Minimum One-Norm Solutions

Instead of the minimum two-norm solution, we can seek a *minimum one-norm* solution, that is, minimize

$$||x||_1 = \sum_{n=1}^{N} |x_n|,$$

subject to $Ax = b$; we denote by x^* the minimum one-norm solution. As we shall see, this problem can be formulated as a linear programming problem, so is easily solved.

The entries of x need not be non-negative, so the problem is not yet a linear programming problem. Let

$$B = \begin{bmatrix} A & -A \end{bmatrix},$$

and consider the linear programming problem of minimizing the function

$$c^T z = \sum_{n=1}^{2N} z_n,$$

subject to the constraints $z \geq 0$, and $Bz = b$. Let z^* be the solution. We write

$$z^* = \begin{bmatrix} u^* \\ v^* \end{bmatrix}.$$

Then, as we shall see, $x^* = u^* - v^*$ minimizes the one-norm, subject to $Ax = b$.

First, we show that $u_n^* v_n^* = 0$, for each n. If this were not the case and there is an n such that $0 < v_n^* < u_n^*$, then we can create a new vector z by replacing the old u_n^* with $u_n^* - v_n^*$ and the old v_n^* with zero, while maintaining $Bz = b$. But then, since $u_n^* - v_n^* < u_n^* + v_n^*$, it follows that $c^T z < c^T z^*$, which is a contradiction. Consequently, we have $\|x^*\|_1 = c^T z^*$.

Now we select any x with $Ax = b$. Write $u_n = x_n$, if $x_n \geq 0$, and $u_n = 0$, otherwise. Let $v_n = u_n - x_n$, so that $x = u - v$. Then let

$$z = \begin{bmatrix} u \\ v \end{bmatrix}.$$

Then $b = Ax = Bz$, and $c^T z = \|x\|_1$. And so,

$$\|x^*\|_1 = c^T z^* \leq c^T z = \|x\|_1,$$

and x^* must be a minimum one-norm solution.

Ex. 7.1 *Find a system of linear equations $Ax = b$ for which there are multiple minimum one-norm solutions.*

7.5 Sparse Solutions

For any vector x, we define the *support* of x to be the subset S of $\{1, 2, ..., N\}$ consisting of those n for which the entries $x_n \neq 0$. For any under-determined system $Ax = b$, there will, of course, be at least one solution, call it x', of minimum support, that is, for which $|S|$, the size of the support set S, is minimum. However, finding such a maximally sparse solution requires combinatorial optimization, and is known to be computationally difficult. It is important, therefore, to have a computationally tractable method for finding maximally sparse solutions.

7.5.1 Maximally Sparse Solutions

Consider the following problem: among all solutions x of the consistent system $Ax = b$, find one, x', that is maximally sparse, that is, has the minimum number of non-zero entries. Obviously, there will be at least one such solution having minimal support, but finding one, however, is a combinatorial optimization problem and is generally NP-hard.

7.5.2 Why the One-Norm?

When a system of linear equations $Ax = b$ is under-determined, we can find the *minimum two-norm solution*. One drawback to this approach is that the two-norm penalizes relatively large values of x_n much more than the smaller ones, so tends to provide non-sparse solutions. Alternatively, we may seek the minimum one-norm solution. The one-norm still penalizes

relatively large entries x_n more than the smaller ones, but much less so than the two-norm does. As a result, it often happens that the minimum one-norm solution actually is a maximally sparse solution, as well.

7.5.3 Comparison with the Weighted Two-Norm Solution

Our intention is to select weights w_n so that w_n^{-1} is reasonably close to the absolute value of the corresponding entry of the minimum one-norm solution $|x_n^*|$; consider, therefore, what happens when $w_n^{-1} = |x_n^*|$. We claim that $\tilde{x}$ is also a minimum-one-norm solution.

To see why this is true, note that, for any x, we have

$$\sum_{n=1}^{N} |x_n| = \sum_{n=1}^{N} \frac{|x_n|}{\sqrt{|x_n^*|}} \sqrt{|x_n^*|}$$

$$\leq \sqrt{\sum_{n=1}^{N} \frac{|x_n|^2}{|x_n^*|}} \sqrt{\sum_{n=1}^{N} |x_n^*|}.$$

Therefore,

$$\sum_{n=1}^{N} |\tilde{x}_n| \leq \sqrt{\sum_{n=1}^{N} \frac{|\tilde{x}_n|^2}{|x_n^*|}} \sqrt{\sum_{n=1}^{N} |x_n^*|}$$

$$\leq \sqrt{\sum_{n=1}^{N} \frac{|x_n^*|^2}{|x_n^*|}} \sqrt{\sum_{n=1}^{N} |x_n^*|} = \sum_{n=1}^{N} |x_n^*|.$$

Therefore, $\tilde{x}$ is also a solution that minimizes the one-norm. If x^* is unique, then $\tilde{x} = x^*$.

7.5.4 Iterative Reweighting

Let x be the truth. generally, we want each weight w_n to be a good prior estimate of the reciprocal of $|x_n|$. Because we do not yet know x, we may take a sequential-optimization approach, beginning with weights $w_n^0 > 0$, finding the minimum weighted two-norm solution using these weights, then using this solution to get a (we hope!) better choice for the weights, and so on. This sequential approach was successfully implemented in the early 1980's by Michael Fiddy and his students [135].

In [74], the same approach is taken, but with respect to the one-norm. Since the one-norm still penalizes larger values disproportionately, balance can be achieved by minimizing a weighted one-norm, with weights close to the reciprocals of the $|x_n|$. Again, not yet knowing x, they employ a sequential approach, using the previous minimum weighted one-norm solution to

obtain the new set of weights for the next minimization. At each step of the sequential procedure, the previous reconstruction is used to estimate the true support of the desired solution.

It is interesting to note that an on-going debate among users of the minimum weighted two-norm approach concerns the nature of the prior weighting. With x denoting the truth, does w_n approximate $|x_n|$ or $|x_n|^2$? This is close to the issue treated in [74], the use of a weight in the minimum one-norm approach.

It should be noted again that finding a sparse solution is not usually the goal in the use of the minimum weighted two-norm approach, but the use of the weights has much the same effect as using the one-norm to find sparse solutions: to the extent that the weights approximate the entries of x^*, their use reduces the penalty associated with the larger entries of an estimated solution.

7.6 Why Sparseness?

One obvious reason for wanting sparse solutions of $Ax = b$ is that we have prior knowledge that the desired solution is sparse. Such a problem arises in signal analysis from Fourier-transform data. In other cases, such as in the reconstruction of locally constant signals, it is not the signal itself, but its discrete derivative, that is sparse.

7.6.1 Signal Analysis

Suppose that our signal $f(t)$ is known to consist of a small number of complex exponentials, so that $f(t)$ has the form

$$f(t) = \sum_{j=1}^{J} a_j e^{i\omega_j t},$$

for some small number of frequencies ω_j in the interval $[0, 2\pi)$. For $n = 0, 1, ..., N-1$, let $f_n = f(n)$, and let f be the N-vector with entries f_n; we assume that J is much smaller than N. The discrete (vector) Fourier transform of f is the vector $\hat{f}$ having the entries

$$\hat{f}_k = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} f_n e^{2\pi ikn/N},$$

for $k = 0, 1, ..., N-1$; we write $\hat{f} = Ef$, where E is the N by N matrix with entries $E_{kn} = \frac{1}{\sqrt{N}} e^{2\pi ikn/N}$. If N is large enough, we may safely assume that

each of the ω_j is equal to one of the frequencies $2\pi ik$ and that the vector $\hat{f}$ is J-sparse. The question now is: How many values of $f(n)$ do we need to calculate in order to be sure that we can recapture $f(t)$ exactly? We have the following theorem [73]:

Theorem 7.1 *Let N be prime. Let S be any subset of $\{0, 1, ..., N-1\}$ with $|S| \geq 2J$. Then the vector $\hat{f}$ can be uniquely determined from the measurements f_n for n in S.*

We know that

$$f = E^\dagger \hat{f},$$

where $E^\dagger$ is the conjugate transpose of the matrix E. The point here is that, for any matrix R obtained from the identity matrix I by deleting $N - |S|$ rows, we can recover the vector $\hat{f}$ from the measurements Rf.

If N is not prime, then the assertion of the theorem may not hold, since we can have $n = 0 \bmod N$, without $n = 0$. However, the assertion remains valid for most sets of J frequencies and most subsets S of indices; therefore, with high probability, we can recover the vector $\hat{f}$ from Rf.

Note that the matrix E is *unitary*, that is, $E^\dagger E = I$, and, equivalently, the columns of E form an orthonormal basis for $\mathbb{C}^N$. The data vector is

$$b = Rf = RE^\dagger \hat{f}.$$

In this example, the vector f is not sparse, but can be represented sparsely in a particular orthonormal basis, namely as $f = E^\dagger \hat{f}$, using a sparse vector $\hat{f}$ of coefficients. The *representing basis* then consists of the columns of the matrix $E^\dagger$. The measurements pertaining to the vector f are the values f_n, for n in S. Since f_n can be viewed as the inner product of f with δ^n, the nth column of the identity matrix I, that is,

$$f_n = \langle \delta^n, f \rangle,$$

the columns of I provide the so-called *sampling basis*. With $A = RE^\dagger$ and $x = \hat{f}$, we then have

$$Ax = b,$$

with the vector x sparse. It is important for what follows to note that the matrix A is random, in the sense that we choose which rows of I to use to form R.

7.6.2 Locally Constant Signals

Suppose now that the function $f(t)$ is locally constant, consisting of some number of horizontal lines. We discretize the function $f(t)$ to get

the vector $f = (f(0), f(1), ..., f(N))^T$. The discrete derivative vector is $g = (g_1, g_2, ..., g_N)^T$, with

$$g_n = f(n) - f(n-1).$$

Since $f(t)$ is locally constant, the vector g is sparse. The data we will have will not typically be values $f(n)$. The goal will be to recover f from M linear functional values pertaining to f, where M is much smaller than N. We shall assume, from now on, that we have measured, or can estimate, the value $f(0)$.

Our M by 1 data vector d consists of measurements pertaining to the vector f:

$$d_m = \sum_{n=0}^{N} H_{mn} f_n,$$

for $m = 1, ..., M$, where the H_{mn} are known. We can then write

$$d_m = f(0)\Big(\sum_{n=0}^{N} H_{mn}\Big) + \sum_{k=1}^{N}\Big(\sum_{j=k}^{N} H_{mj}\Big) g_k.$$

Since $f(0)$ is known, we can write

$$b_m = d_m - f(0)\Big(\sum_{n=0}^{N} H_{mn}\Big) = \sum_{k=1}^{N} A_{mk} g_k,$$

where

$$A_{mk} = \sum_{j=k}^{N} H_{mj}.$$

The problem is then to find a sparse solution of $Ax = g$. As in the previous example, we often have the freedom to select the linear functions, that is, the values H_{mn}, so the matrix A can be viewed as random.

7.6.3 Tomographic Imaging

The reconstruction of tomographic images is an important aspect of medical diagnosis, and one that combines aspects of both of the previous examples. The data one obtains from the scanning process can often be interpreted as values of the Fourier transform of the desired image; this is precisely the case in magnetic-resonance imaging, and approximately true for x-ray transmission tomography, positron-emission tomography (PET) and single-photon emission tomography (SPECT). The images one encounters in medical diagnosis are often approximately locally constant, so the associated array of discrete partial derivatives will be sparse. If this sparse derivative array can be recovered from relatively few Fourier-transform values, then the scanning time can be reduced.

7.7 Positive Linear Systems

When the entries of the matrix A are non-negative, the entries of the vector b are positive, and we require that the entries of x be non-negative, we say that we have a *positive system*. We call the system *under-determined* when there are multiple non-negative solutions. It is appropriate now to use the cross-entropy, or Kullback-Leibler (KL), distance between non-negative vectors, rather than the two-norm or the one-norm.

In the under-determined case, the MART and its block-iterative versions, the RBI-SMART algorithms, all converge to that non-negative solution x for which $KL(x, x^0)$ is minimized. The EMML algorithm and its block-iterative variants also converge to non-negative solutions, but they may not all be the same solution, and no explicit characterization of these solutions is known; that is, they depend on x^0, but precisely how is not known.

When we wish to impose further constraints on the entries of x, we can use the ABMART or the ABEMML algorithms.

7.8 Feasible-Point Methods

In previous sections we considered the minimum two-norm and minimum one-norm solutions for under-determined systems $Ax = b$. A more general approach is to minimize some function $f(x)$, subject to $Ax = b$, which is the subject of this section.

We consider now the problem of minimizing the function $f(x) : \mathbb{R}^N \rightarrow \mathbb{R}$, subject to the equality constraints $Ax = b$, where A is an M by N real matrix, with rank M and $M < N$. The two methods we consider here are *feasible-point methods*, also called *interior-point methods*.

7.8.1 The Reduced Newton-Raphson Method

The first method we consider is a modification of the Newton-Raphson method, in which we begin with a feasible point and each NR step is projected into the null space of the matrix A, to maintain the condition $Ax = b$. The discussion here is taken from [209].

Let $\hat{x}$ be a *feasible point*, that is, $A\hat{x} = b$. Then $x = \hat{x} + p$ is also feasible if p is in the null space of A, that is, $Ap = 0$. Let Z be an N by $N - M$ matrix whose columns form a basis for the null space of A. We want $p = Zv$

for some v. The best v will be the one for which the function

$$\phi(v) = f(\hat{x} + Zv)$$

is minimized. We can apply to the function $\phi(v)$ the steepest descent method, or Newton-Raphson or any other minimization technique. The steepest descent method, applied to $\phi(v)$, is called the *reduced steepest descent method*; the Newton-Raphson method, applied to $\phi(v)$, is called the *reduced Newton-Raphson method*. The gradient of $\phi(v)$, also called the *reduced gradient*, is

$$\nabla\phi(v) = Z^T \nabla f(x),$$

and the Hessian matrix of $\phi(v)$, also called the *reduced Hessian matrix*, is

$$\nabla^2\phi(v) = Z^T \nabla^2 f(x) Z,$$

where $x = \hat{x} + Zv$, so algorithms to minimize $\phi(v)$ can be written in terms of the gradient and Hessian of f itself.

The reduced NR algorithm can then be viewed in terms of the vectors $\{v^k\}$, with $v^0 = 0$ and

$$v^{k+1} = v^k - [\nabla^2\phi(v^k)]^{-1}\nabla\phi(v^k); \tag{7.1}$$

the corresponding x^k is

$$x^k = \hat{x} + Zv^k.$$

7.8.1.1 An Example

Consider the problem of minimizing the function

$$f(x) = \frac{1}{2}x_1^2 - \frac{1}{2}x_3^2 + 4x_1x_2 + 3x_1x_3 - 2x_2x_3,$$

subject to

$$x_1 - x_2 - x_3 = -1.$$

Let $\hat{x} = [1, 1, 1]^T$. Then the matrix A is $A = [1, -1, -1]$ and the vector b is $b = [-1]$. Let the matrix Z be

$$Z = \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}. \tag{7.2}$$

The reduced gradient at $\hat{x}$ is then

$$Z^T \nabla f(\hat{x}) = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 8 \\ 2 \\ 0 \end{bmatrix} = \begin{bmatrix} 10 \\ 8 \end{bmatrix}, \tag{7.3}$$

and the reduced Hessian matrix at $\hat{x}$ is

$$Z^T \nabla^2 f(\hat{x}) Z = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 4 & 3 \\ 4 & 0 & -2 \\ 3 & -2 & -1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 9 & 6 \\ 6 & 6 \end{bmatrix}. \tag{7.4}$$

Then the reduced Newton-Raphson equation yields

$$v = \begin{bmatrix} -2/3 \\ -2/3 \end{bmatrix}, \tag{7.5}$$

and the reduced Newton-Raphson direction is

$$p = Zv = \begin{bmatrix} -4/3 \\ -2/3 \\ -2/3 \end{bmatrix}. \tag{7.6}$$

Since the function $\phi(v)$ is quadratic, one reduced Newton-Raphson step suffices to obtain the solution, $x^* = [-1/3, 1/3, 1/3]^T$.

7.8.2 A Primal-Dual Approach

Once again, the objective is to minimize the function $f(x) : \mathbb{R}^N \to \mathbb{R}$, subject to the equality constraints $Ax = b$. According to the Karush-Kuhn-Tucker Theorem [70], $\nabla L(x, \lambda) = 0$ at the optimal values of x and λ, where the Lagrangian $L(x, \lambda)$ is

$$L(x, \lambda) = f(x) + \lambda^T (b - Ax).$$

Finding a zero of the gradient of $L(x, \lambda)$ means that we have to solve the equations

$$\nabla f(x) - A^T \lambda = 0$$

and

$$Ax = b.$$

We define the function $G(x, \lambda)$ taking values in $\mathbb{R}^N \times \mathbb{R}^M$ to be

$$G(x, \lambda) = (\nabla f(x) - A^T \lambda, Ax - b)^T.$$

We then apply the NR method to find a zero of the function G. The Jacobian matrix for G is

$$J_G(x, \lambda) = \begin{bmatrix} \nabla^2 f(x) & -A^T \\ A & 0 \end{bmatrix},$$

so one step of the NR method is

$$(x^{k+1}, \lambda^{k+1})^T = (x^k, \lambda^k)^T - J_G(x^k, \lambda^k)^{-1} G(x^k, \lambda^k). \tag{7.7}$$

We can rewrite this as

$$\nabla^2 f(x^k)(x^{k+1} - x^k) - A^T(\lambda^{k+1} - \lambda^k) = A^T\lambda^k - \nabla f(x^k), \tag{7.8}$$

and

$$A(x^{k+1} - x^k) = b - Ax^k. \tag{7.9}$$

It follows from Equation (7.9) that $Ax^{k+1} = b$, for $k = 0, 1, \ldots$, so that this primal-dual algorithm is a feasible-point algorithm.

Chapter 8

Convex Sets

8.1 Chapter Summary

Convex sets and convex functions play important roles in linear algebra and optimization. Subspaces, hyperplanes and half-spaces are convex sets that are naturally associated with linear equations and inequalities. In this chapter we survey the basic facts concerning the geometry of convex sets.

8.2 A Bit of Topology

Having a norm allows us to define the distance between two points x and y in $\mathbb{R}^J$ as $||x-y||$. Being able to talk about how close points are to each other enables us to define continuity of functions on $\mathbb{R}^J$ and to consider topological notions of closed set, open set, interior of a set and boundary of a set. While the actual numerical distance between points will vary as we change norm, all of these purely topological notions are independent of the particular norm being used. When we are required to limit our discussion to the two-norm we shall indicate that by using the notation $\|x\|_2$ instead of the general $\|x\|$. Whenever we speak of "orthogonality" , we are in the context of an inner product space and the norm will be the two-norm derived from the inner product.

Definition 8.1 *A subset B of $\mathbb{R}^J$ is* closed *if, whenever x^k is in B for*

each non-negative integer k and $||x - x^k|| \to 0$, as $k \to +\infty$, then x is in B.

For example, $B = [0, 1]$ is closed as a subset of $\mathbb{R}$, but $B = (0, 1)$ is not.

Definition 8.2 *We say that $d \geq 0$ is the* distance from the point x *to the set B if, for every $\epsilon > 0$, there is b_ϵ in B, with $||x - b_\epsilon|| < d + \epsilon$, and no b in B with $||x - b|| < d$.*

The Euclidean distance from the point 0 in $\mathbb{R}$ to the set $(0, 1)$ is zero, while its distance to the set $(1, 2)$ is one. It follows easily from the definitions that, if B is closed and $d = 0$, then x is in B.

Definition 8.3 *The* closure *of a set B is the set of all points x whose distance from B is zero.*

The closure of the interval $B = (0, 1)$ is $[0, 1]$.

Definition 8.4 *A subset U of $\mathbb{R}^J$ is* open *if its complement, the set of all points not in U, is closed.*

Definition 8.5 *Let C be a subset of $\mathbb{R}^J$. A point x in C is said to be an* interior point *of set C if there is $\epsilon > 0$ such that every point z with $||x - z|| < \epsilon$ is in C. The* interior *of the set C, written int(C), is the set of all interior points of C. It is also the largest open set contained within C.*

For example, the open interval $(0, 1)$ is the interior of the intervals $(0, 1]$ and $[0, 1]$. A set C is open if and only if $C = \text{int}(C)$.

Definition 8.6 *A point x in $\mathbb{R}^J$ is said to be a* boundary point *of set C if, for every $\epsilon > 0$, there are points y_ϵ in C and z_ϵ not in C, both depending on the choice of ϵ, with $||x - y_\epsilon|| < \epsilon$ and $||x - z_\epsilon|| < \epsilon$. The* boundary *of C is the set of all boundary points of C. It is also the intersection of the closure of C with the closure of its complement.*

For example, the points $x = 0$ and $x = 1$ are boundary points of the set $(0, 1]$.

Definition 8.7 *For $k = 0, 1, 2, ...$, let x^k be a vector in $\mathbb{R}^J$. The sequence of vectors $\{x^k\}$ is said to* converge *to the vector z if, given any $\epsilon > 0$, there is positive integer n, usually depending on ϵ, such that, for every $k > n$, we have $||z - x^k|| \leq \epsilon$. Then we say that z is the* limit of the sequence.

For example, the sequence $\{x^k = \frac{1}{k+1}\}$ in $\mathbb{R}$ converges to $z = 0$. The sequence $\{(-1)^k\}$ alternates between 1 and -1, so does not converge. However, the subsequence associated with odd k converges to $z = -1$, while the subsequence associated with even k converges to $z = 1$. The values $z = -1$ and $z = 1$ are called *subsequential limit points*, or, sometimes, *cluster points* of the sequence.

Definition 8.8 *A sequence $\{x^k\}$ of vectors in $\mathbb{R}^J$ is said to be* bounded *if there is a constant $b > 0$, such that $||x^k|| \leq b$, for all k.*

A fundamental result in analysis is the following.

Proposition 8.1 *Every convergent sequence of vectors in $\mathbb{R}^J$ is bounded. Every bounded sequence of vectors in $\mathbb{R}^J$ has at least one convergent subsequence, therefore, has at least one cluster point.*

8.3 Convex Sets in $\mathbb{R}^J$

In preparation for our discussion of orthogonal projection operators, we consider some of the basic concepts from the geometry of convex sets.

8.3.1 Basic Definitions

We begin with the basic definitions.

Definition 8.9 *A vector z is said to be a* convex combination *of the vectors x and y if there is α in the interval $[0, 1]$ such that $z = (1 - \alpha)x + \alpha y$.*

Definition 8.10 *A nonempty set C in $\mathbb{R}^J$ is said to be* convex *if, for any distinct points x and y in C, and for any real number α in the interval $(0, 1)$, the point $(1 - \alpha)x + \alpha y$ is also in C; that is, C is closed to convex combinations.*

For example, the two-norm unit ball B in $\mathbb{R}^J$, consisting of all x with $||x||_2 \leq 1$, is convex, while the surface of the ball, the set of all x with $||x||_2 = 1$, is not convex. In fact, the unit ball in any norm is a closed convex set.

Definition 8.11 *The* convex hull *of a set S, denoted conv(S), is the smallest convex set containing S.*

Proposition 8.2 *The convex hull of a set S is the set C of all convex combinations of members of S.*

Definition 8.12 *A subset S of $\mathbb{R}^J$ is a* subspace *if, for every x and y in S and scalars α and β, the linear combination $\alpha x + \beta y$ is again in S.*

A subspace is necessarily a convex set.

Definition 8.13 *The* orthogonal complement *of a subspace S is the set*

$$S^{\perp} = \{u | u^T s = 0, \text{for every } s \in S\}, \tag{8.1}$$

the set of all vectors u in $\mathbb{R}^J$ that are orthogonal to every member of S.

For example, in $\mathbb{R}^3$, the x, y-plane is a subspace and has for its orthogonal complement the z-axis.

Definition 8.14 *A subset M of $\mathbb{R}^J$ is a* linear manifold *if there is a subspace S and a vector b such that*

$$M = S + b = \{x | x = s + b, \text{for some } s \text{ in } S\}.$$

Any linear manifold is convex.

Definition 8.15 *For a fixed column vector a with Euclidean length one and a fixed scalar γ the* hyperplane *determined by a and γ is the set*

$$H(a, \gamma) = \{z | \langle a, z \rangle = \gamma\}.$$

The hyperplanes $H(a, \gamma)$ are linear manifolds, and the hyperplanes $H(a, 0)$ are subspaces.

Definition 8.16 *Given a subset C of $\mathbb{R}^J$, the* affine hull *of C, denoted aff(C), is the smallest linear manifold containing C.*

For example, let C be the line segment connecting the two points $(0, 1)$ and $(1, 2)$ in $\mathbb{R}^2$. The affine hull of C is the straight line whose equation is $y = x + 1$.

Definition 8.17 *The* dimension *of a subset of $\mathbb{R}^J$ is the dimension of its affine hull, which is the dimension of the subspace of which it is a translate.*

The set C above has dimension one. A set containing only one point is its own affine hull, since it is a translate of the subspace $\{0\}$.

In $\mathbb{R}^2$, the line segment connecting the points $(0, 1)$ and $(1, 2)$ has no interior; it is a one-dimensional subset of a two-dimensional space and can contain no two-dimensional ball. But, the part of this set without its two end points is a sort of interior, called the *relative interior*.

Definition 8.18 *The* relative interior *of a subset C of $\mathbb{R}^J$, denoted ri(C), is the interior of C, as defined by considering C as a subset of its affine hull.*

Since a set consisting of a single point is its own affine hull, it is its own relative interior.

Definition 8.19 *A point x in a convex set C is said to be an* extreme point *of C if the set obtained by removing x from C remains convex.*

Said another way, $x \in C$ is an extreme point of C if x cannot be written as

$$x = (1 - \alpha)y + \alpha z, \tag{8.2}$$

for $y, z \neq x$ and $\alpha \in (0, 1)$. For example, the point $x = 1$ is an extreme point of the convex set $C = [0, 1]$. Every point on the boundary of a sphere in $\mathbb{R}^J$ is an extreme point of the sphere. The set of all extreme points of a convex set is denoted $\text{Ext}(C)$.

Definition 8.20 *A non-zero vector d is said to be a* direction of unboundedness *of a convex set C if, for all x in C and all $\gamma \geq 0$, the vector $x + \gamma d$ is in C.*

For example, if C is the non-negative orthant in $\mathbb{R}^J$, then any non-negative vector d is a direction of unboundedness.

Definition 8.21 *A vector a is* normal *to a convex set C at the point s in C if*

$$\langle a, c - s \rangle \leq 0, \tag{8.3}$$

for all c in C.

Definition 8.22 *Let C be convex and s in C. The* normal cone *to C at s, denoted $N_C(s)$, is the set of all vectors a that are normal to C at s.*

The notions of normality and normal cone make sense only in the presence of an inner product, and therefore, will be discussed only when the norm is the two-norm.

8.3.2 Orthogonal Projection onto Convex Sets

The following proposition is fundamental in the study of convexity and can be found in most books on the subject; see, for example, the text by Goebel and Reich [147].

Proposition 8.3 *Given any nonempty closed convex set C and an arbitrary vector x in $\mathbb{R}^J$, there is a unique member $P_C x$ of C closest to x, in the two-norm. The vector $P_C x$ is called the orthogonal (or metric) projection of x onto C, and the operator P_C is the orthogonal projection onto C.*

Proof: If x is in C, then $P_C x = x$, so assume that x is not in C. Then $d > 0$, where d is the distance from x to C. For each positive integer n, select c^n in C with $||x - c^n||_2 < d + \frac{1}{n}$. Then, since for all n we have

$$\|c^n\|_2 = \|c^n - x + x\|_2 \leq \|c^n - x\|_2 + \|x\|_2 \leq d + \frac{1}{n} + \|x\|_2 < d + 1 + \|x\|_2,$$

the sequence $\{c^n\}$ is bounded; let c^* be any cluster point. It follows easily that $||x - c^*||_2 = d$ and that c^* is in C. If there is any other member c of C with $||x - c||_2 = d$, then, by the Parallelogram Law, we would have $||x - (c^* + c)/2||_2 < d$, which is a contradiction. Therefore, c^* is $P_C x$. ∎

For example, if $C = U$, the unit ball, then $P_C x = x/||x||_2$, for all x such that $||x||_2 > 1$, and $P_C x = x$ otherwise. If C is $\mathbb{R}^J_+$, the nonnegative cone of $\mathbb{R}^J$, consisting of all vectors x with $x_j \geq 0$, for each j, then $P_C x = x_+$, the vector whose entries are $\max(x_j, 0)$. For any closed, convex set C, the distance from x to C is $||x - P_C x||_2$.

If a nonempty set S is not convex, then the orthogonal projection of a vector x onto S need not be well defined; there may be more than one vector in S closest to x. In fact, it is known that a set S is convex if and only if, for every x not in S, there is a unique point in S closest to x. Note that there may well be some x for which there is a unique closest point in S, but if S is not convex, then there must be at least one point without a unique closest point in S.

The main reason for not speaking about orthogonal projection onto convex sets in other norms is that the nearest point to x in C need not be unique; remember, the parallelogram law need not hold. For example, consider the closed convex set C in $\mathbb{R}^2$ consisting of all vectors $(a, b)^T$ with $a \geq 0$, $b \geq 0$, and $a + b = 1$. Let $x = (1, 1)^T$. Then each point in C is a distance one from x, in the sense of the one-norm.

Lemma 8.1 *For $H = H(a, \gamma)$, $z = P_H x$ is the vector*

$$z = P_H x = x + (\gamma - \langle a, x \rangle)a. \tag{8.4}$$

We shall use this fact in our discussion of the ART algorithm.

For an arbitrary nonempty closed convex set C in $\mathbb{R}^J$, the orthogonal projection $T = P_C$ is a nonlinear operator, unless, of course, C is a subspace. We may not be able to describe $P_C x$ explicitly, but we do know a useful property of $P_C x$.

Proposition 8.4 *For a given x, a vector z in C is $P_C x$ if and only if*

$$\langle c - z, z - x \rangle \geq 0, \tag{8.5}$$

for all c in the set C.

Proof: Let c be arbitrary in C and α in $(0,1)$. Then

$$||x - P_C x||_2^2 \leq ||x - (1-\alpha)P_C x - \alpha c||_2^2 = ||x - P_C x + \alpha(P_C x - c)||_2^2$$

$$= ||x - P_C x||_2^2 - 2\alpha\langle x - P_C x, c - P_C x\rangle + \alpha^2||P_C x - c||_2^2. \tag{8.6}$$

Therefore,

$$-2\alpha\langle x - P_C x, c - P_C x\rangle + \alpha^2||P_C x - c||_2^2 \geq 0, \tag{8.7}$$

so that

$$2\langle x - P_C x, c - P_C x\rangle \leq \alpha||P_C x - c||_2^2. \tag{8.8}$$

Taking the limit, as $\alpha \to 0$, we conclude that

$$\langle c - P_C x, P_C x - x\rangle \geq 0. \tag{8.9}$$

If z is a member of C that also has the property

$$\langle c - z, z - x\rangle \geq 0, \tag{8.10}$$

for all c in C, then we have both

$$\langle z - P_C x, P_C x - x\rangle \geq 0, \tag{8.11}$$

and

$$\langle z - P_C x, x - z\rangle \geq 0. \tag{8.12}$$

Adding on both sides of these two inequalities lead to

$$\langle z - P_C x, P_C x - z\rangle \geq 0. \tag{8.13}$$

But,

$$\langle z - P_C x, P_C x - z\rangle = -||z - P_C x||_2^2, \tag{8.14}$$

so it must be the case that $z = P_C x$. This completes the proof. ∎

8.4 Geometric Interpretations of $\mathbb{R}^J$

The word *vector* is used in two distinct ways when we speak of a member of $\mathbb{R}^J$ and this can lead to some confusion. When we say that x is a member

of $\mathbb{R}^J$ we mean that x is a J by 1 real matrix, that is, a single column, with J real entries x_j. By choosing to make x a column and not a row we make it possible to write Ax, where A is a matrix with J columns. One linear equation in J variables has the form

$$a_1x_1 + a_2x_2 + ... + a_Jx_J = \gamma. \tag{8.15}$$

We denote by a the J by 1 matrix with entries a_j, so that Equation (8.15) can be written in the compact form $a^Tx = \gamma$. Because the dot product $a \cdot x$ of a and x is also the standard inner product on $\mathbb{R}^J$, we have these equivalent ways of writing Equation (8.15):

$$a \cdot x = \langle a, x \rangle = a^Tx = \gamma. \tag{8.16}$$

We commonly visualize the space $\mathbb{R}^J$ by thinking of $\mathbb{R}^2$ or $\mathbb{R}^3$, and interpreting the members x of $\mathbb{R}^J$ as specific locations or points in J-dimensional space. We think of $x = (3, -4)^T$ in $\mathbb{R}^2$ as the point we reach from the origin by going three units "east" and four units "south" .

In any inner product space V it is natural to speak about *orthogonality* and to say that members u and v of V are *orthogonal* if $\langle u, v \rangle = 0$. This can lead to some confusion when we seek a geometric interpretation of $\mathbb{R}^J$.

When we discuss the plane $\mathbb{R}^2$ or three-dimensional space $\mathbb{R}^3$, we also use the word *vector* to mean a directed line segment. For the purposes of this section, and nowhere else in this text, we shall denote a directed line segment by $\mathbf{x} = (x_1, x_2)^T$. It is unfortunate that this notational distinction is not maintained in most literature on the subject; for that reason, we do not maintain it in this text. It is left to the reader to infer the proper interpretation from the context.

The directed line segment $\mathbf{x}$ and the vector $x = (x_1, x_2)^T$ in $\mathbb{R}^2$ are obviously related. The directed line segment $\mathbf{x}$ is what we obtain if we begin the segment at $0 = (0, 0)^T$ and end at $x = (x_1, x_2)^T$. But $\mathbf{x}$ is also what we get if we begin the segment at $(-2, 7)^T$ and end it at $(-2 + x_1, 7 + x_2)^T$. Vectors $\mathbf{x}$ interpreted as directed line segments in $\mathbb{R}^J$ have no fixed location in $\mathbb{R}^J$, unlike vectors x interpreted as points in $\mathbb{R}^J$.

When we consider a single linear equation, as in Equation (8.15), it is convenient to define the *hyperplane* $H(a, \gamma)$ in $\mathbb{R}^J$ associated with this equation as

$$H(a, \gamma) = \{x | a \cdot x = a^Tx = \langle a, x \rangle = \gamma\}. \tag{8.17}$$

Note that the symbol $x|$ means "all x for which the following property holds" .

The members of the hyperplane are to be thought of as points in $\mathbb{R}^J$ having a specific location, not as directed line segments. On the other hand, it is helpful to think of a as $\mathbf{a}$, a directed line segment, so that the first

term in $\langle a, x\rangle$ is a directed line segment, while the second term is a point with a fixed location. If x and z are any two members of $H(a, \gamma)$, then the directed line segment that goes from z to x, which is $\mathbf{x} - \mathbf{z}$, is orthogonal to the directed line segment $\mathbf{a}$. We say then that $\mathbf{a}$ is a vector that is *normal* to $H(a, \gamma)$. Note that we do not mean that $a \cdot x = 0$ for each x in $H(a, \gamma)$, but that $a \cdot (x - z) = 0$ for all x and z in $H(a, \gamma)$.

8.5 Some Results on Projections

The characterization of the orthogonal projection operator P_C given by Proposition 8.4 has a number of important consequences.

Corollary 8.1 *Let S be any subspace of $\mathbb{R}^J$. Then, for any x in $\mathbb{R}^J$ and s in S, we have*

$$\langle P_S x - x, s\rangle = 0. \tag{8.18}$$

Proof: Since S is a subspace, $s + P_S x$ is again in S, for all s, as is cs, for every scalar c. ∎

This corollary enables us to prove the Decomposition Theorem.

Theorem 8.1 *Let S be any subspace of $\mathbb{R}^J$ and x any member of $\mathbb{R}^J$. Then there are unique vectors s in S and u in $S^\perp$ such that $x = s + u$. The vector s is $P_S x$ and the vector u is $P_{S^\perp} x$.*

Proof: For the given x we take $s = P_S x$ and $u = x - P_S x$. Corollary 8.1 assures us that u is in $S^\perp$. Now we need to show that this decomposition is unique. To that end, suppose that we can write $x = s_1 + u_1$, with s_1 in S and u_1 in $S^\perp$. Then Proposition 8.4 tells us that, since $s_1 - x$ is orthogonal to every member of S, s_1 must be $P_S x$. ∎

This theorem is often presented in a slightly different manner.

Theorem 8.2 *Let A be a real I by J matrix. Then every vector b in $\mathbb{R}^I$ can be written uniquely as $b = Ax + w$, where $A^T w = 0$.*

To derive Theorem 8.2 from Theorem 8.1, we simply let $S = \{Ax | x \in \mathbb{R}^J\}$. Then $S^\perp$ is the set of all w such that $A^T w = 0$. It follows that w is the member of the null space of A^T closest to b.

Here are additional consequences of Proposition 8.4.

Corollary 8.2 *Let S be any subspace of $\mathbb{R}^J$, d a fixed vector, and V the linear manifold $V = S + d = \{v = s + d | s \in S\}$, obtained by translating the members of S by the vector d. Then, for every x in $\mathbb{R}^J$ and every v in V, we have*

$$\langle P_V x - x, v - P_V x \rangle = 0. \tag{8.19}$$

Proof: Since v and $P_V x$ are in V, they have the form $v = s + d$, and $P_V x = \hat{s} + d$, for some s and $\hat{s}$ in S. Then $v - P_V x = s - \hat{s}$. ∎

Corollary 8.3 *Let H be the hyperplane $H(a, \gamma)$. Then, for every x, and every h in H, we have*

$$\langle P_H x - x, h - P_H x \rangle = 0. \tag{8.20}$$

Corollary 8.4 *Let S be a subspace of $\mathbb{R}^J$. Then $(S^\perp)^\perp = S$.*

Proof: Every x in $\mathbb{R}^J$ has the form $x = s + u$, with s in S and u in $S^\perp$. Suppose x is in $(S^\perp)^\perp$. Then $u = 0$. ∎

Theorem 8.3 (The Separation Theorem) *Let C b e a non-empty, closed convex set in $\mathbb{R}^J$ and x not in C. Then there exists a vector a in $\mathbb{R}^J$ and a scalar γ such that, for all c in C,*

$$\langle a, x \rangle < \gamma \leq \langle a, c \rangle.$$

Proof: Let $a = P_C x - x$. Then, using the inequality

$$\langle P_C x - x, c - P_C x \rangle \geq 0,$$

for all c in C, and the fact that the positive distance d from x to C is $d = \|P_C x - x\|$, we have

$$\langle a, c \rangle \geq \gamma = \langle a, P_C x \rangle = \langle a, P_C x - x \rangle + \langle a, x \rangle = d^2 + \langle a, x \rangle.$$

∎

Chapter 9

Linear Inequalities

9.1 Chapter Summary

Most books on linear algebra devote considerable space to the problem of solving a consistent or inconsistent system of linear equations, say $Ax = b$. Problems involving linear inequalities, such as solving $Ax \geq b$, attract less attention, although such problems play a crucial role in linear programming. The term *linear programming* (LP) refers to the problem of optimizing a linear function of several variables over linear equality or inequality constraints. Such problems arise in many areas of applications. In linear programming, the *primary problem in standard form* (PS) is to minimize $c^T x$, subject to $Ax = b$ and $x \geq 0$. We can view this as one approach to solving an under-determined system, $Ax = b$, by imposing additional constraints on the solution x. It is common, in applications, for A to be quite large, necessitating the use of an iterative algorithm to solve the problem. Dantzig's *Simplex Method* (see [70]) is the best known iterative method for solving LP problems.

9.2 Theorems of the Alternative

Later in this chapter we shall present David Gale's proof of his *strong duality theorem* in linear programming ([141]). His proof makes use of a theorem concerning linear inequalities known as a *theorem of the alternative*. For that reason, we begin with a discussion of these types of theorems.

9.2.1 A Theorem of the Alternative

The following theorem is a good illustration of a type of theorem known as *Theorems of the Alternative*. These theorems assert that precisely one of two problems will have a solution. The proof illustrates how we should go about proving such theorems.

Theorem 9.1 (Gale I)[141] *Precisely one of the following is true:*

- **(1)** *there is x such that $Ax = b$;*
- **(2)** *there is y such that $A^T y = 0$ and $b^T y = 1$.*

Proof: First, we show that it is not possible for both to be true at the same time. Suppose that $Ax = b$ and $A^T y = 0$. Then $b^T y = x^T A^T y = 0$, so that we cannot have $b^T y = 1$. By Theorem 8.1, the fundamental decomposition theorem from linear algebra, we know that, for any b, there are unique Ax and w with $A^T w = 0$ such that $b = Ax + w$. Clearly, $b = Ax$ if and only if $w = 0$. Also, $b^T y = w^T y$. Therefore, if alternative **(1)** does not hold, we must have w non-zero, in which case $A^T y = 0$ and $b^T y = 1$, for $y = w/||w||^2$, so alternative **(2)** holds. ∎

In this section we consider several other theorems of this type.

9.2.2 More Theorems of the Alternative

Theorem 9.2 (Farkas' Lemma)[128] *Precisely one of the following is true:*

- **(1)** *there is $x \geq 0$ such that $Ax = b$;*
- **(2)** *there is y such that $A^T y \geq 0$ and $b^T y < 0$.*

Proof: We can restate the lemma as follows: there is a vector y with $A^T y \geq 0$ and $b^T y < 0$ if and only if b is not a member of the convex set

$C = \{Ax | x \geq 0\}$. If b is not in C, which is closed and convex, then, by the Separation Theorem 8.3, there is a non-zero vector a and real α with

$$a^T b < \alpha \leq a^T Ax = (A^T a)^T x,$$

for all $x \geq 0$. Since $(A^T a)^T x$ is bounded below, as x runs over all non-negative vectors, it follows that $A^T a \geq 0$. Choosing $x = 0$, we have $\alpha \leq 0$. Then let $y = a$. Conversely, if $Ax = b$ does have a non-negative solution x, then $A^T y \geq 0$ implies that $0 \leq y^T Ax = y^T b \geq 0$. ∎

The next theorem can be obtained from Farkas' Lemma.

Theorem 9.3 (Gale II)[141] *Precisely one of the following is true:*

- **(1)** *there is x such that $Ax \leq b$;*
- **(2)** *there is $y \geq 0$ such that $A^T y = 0$ and $b^T y < 0$.*

Proof: First, if both are true, then $0 \leq y^T(b - Ax) = y^T b - 0 = y^T b$, which is a contradiction. Now assume that **(2)** does not hold. Therefore, for every $y \geq 0$ with $A^T y = 0$, we have $b^T y \geq 0$. Let $B = \begin{bmatrix} A & b \end{bmatrix}$. Then the system $B^T y = \begin{bmatrix} 0^T & -1 \end{bmatrix}^T$ has no non-negative solution. Applying Farkas' Lemma, we find that there is a vector $w = \begin{bmatrix} z^T & \gamma \end{bmatrix}^T$ with $Bw \geq 0$ and $\begin{bmatrix} 0^T & -1 \end{bmatrix} w < 0$. So, $Az + \gamma b \geq 0$ and $\gamma > 0$. Let $x = -\frac{1}{\gamma} z$ to get $Ax \leq b$, so that **(1)** holds. ∎

Theorem 9.4 (Gordan)[150] *Precisely one of the following is true:*

- **(1)** *there is x such that $Ax < 0$;*
- **(2)** *there is $y \geq 0$, $y \neq 0$, such that $A^T y = 0$.*

Proof: First, if both are true, then $0 < -y^T Ax = 0$, which cannot be true. Now assume that there is no non-zero $y \geq 0$ with $A^T y = 0$. Then, with $e = (1, 1, ..., 1)^T$, $C = \begin{bmatrix} A & e \end{bmatrix}$, and $d = (0, 0, ..., 0, 1)^T$, there is no non-negative solution of $C^T y = d$. From Farkas' Lemma we then know that there is a vector $z = \begin{bmatrix} u^T & \gamma \end{bmatrix}^T$, with $Cz = Au + \gamma e \geq 0$, and $d^T z < 0$. Then $Ax < 0$ for $x = -u$. ∎

Here are several more theorems of the alternative.

Theorem 9.5 (Stiemke I)[245] *Precisely one of the following is true:*

- **(1)** *there is x such that $Ax \leq 0$ and $Ax \neq 0$;*
- **(2)** *there is $y > 0$ such that $A^T y = 0$.*

Theorem 9.6 (Stiemke II)[245] *Let c be a fixed non-zero vector. Precisely one of the following is true:*

- **(1)** *there is* x *such that* $Ax \leq 0$ *and* $c^T x \geq 0$ *and not both* $Ax = 0$ *and* $c^T x = 0$*;*
- **(2)** *there is* $y > 0$ *such that* $A^T y = c$*.*

Theorem 9.7 (Gale III)[141] *Let* c *be a fixed non-zero vector. Precisely one of the following is true:*

- **(1)** *there is* $x \geq 0$ *such that* $Ax \geq 0$ *and* $c^T x < 0$*;*
- **(2)** *there is* $y \geq 0$ *such that* $A^T y \leq c$*.*

Proof: First, note that we cannot have both true at the same time, since we would then have

$$0 \leq x^T(c - A^T y) = c^T x - (Ax)^T y \leq c^T x < 0,$$

which is a contradiction. Now suppose that **(2)** does not hold. Then there is no $w = [y^T \ v^T]^T \geq 0$ such that

$$\begin{bmatrix} A^T & I \end{bmatrix} w = A^T y + v = c.$$

By Farkas' Lemma (Theorem 9.2), it follows that there is x with

$$\begin{bmatrix} A \\ I \end{bmatrix} x \geq 0,$$

and $c^T x < 0$. Therefore, $Ax \geq 0$, $Ix = x \geq 0$, and $c^T x < 0$; therefore, **(1)** holds. ∎

Theorem 9.8 (Von Neumann)[259] *Precisely one of the following is true:*

- **(1)** *there is* $x \geq 0$ *such that* $Ax > 0$*;*
- **(2)** *there is* $y \geq 0$*,* $y \neq 0$*, such that* $A^T y \leq 0$*.*

Proof: If both were true, then we would have

$$0 < (Ax)^T y = x^T(A^T y),$$

so that $A^T y \leq 0$ would be false. Now suppose that **(2)** does not hold. Then there is no $y \geq 0$, $y \neq 0$, with $A^T y \leq 0$. Consequently, there is no $y \geq 0$, $y \neq 0$, such that

$$\begin{bmatrix} A^T \\ -u^T \end{bmatrix} y = \begin{bmatrix} A^T y \\ -u^T y \end{bmatrix} \leq \begin{bmatrix} 0 \\ -1 \end{bmatrix},$$

where $u^T = (1, 1, ..., 1)$. By Theorem 9.7, there is

$$z = \begin{bmatrix} x \\ \alpha \end{bmatrix} \geq 0,$$

such that

$$\begin{bmatrix} A & -u \end{bmatrix} z = \begin{bmatrix} A & -u \end{bmatrix} \begin{bmatrix} x \\ \alpha \end{bmatrix} \geq 0,$$

and

$$\begin{bmatrix} 0^T & -1 \end{bmatrix} z = \begin{bmatrix} 0^T & -1 \end{bmatrix} \begin{bmatrix} x \\ \alpha \end{bmatrix} = -\alpha < 0.$$

Therefore, $\alpha > 0$ and $(Ax)_i - \alpha \geq 0$ for each i, and so $Ax > 0$ and **(1)** holds. ∎

Theorem 9.9 (Tucker)[251] *Precisely one of the following is true:*

- **(1)** *there is* $x \geq 0$ *such that* $Ax \geq 0$, $Ax \neq 0$*;*
- **(2)** *there is* $y > 0$ *such that* $A^T y \leq 0$.

9.2.3 Another Proof of Farkas' Lemma

In the previous section, we proved Farkas' Lemma, Theorem 9.2, using the Separation Theorem, the proof of which, in turn, depended here on the existence of the orthogonal projection onto any closed convex set. It is possible to prove Farkas' Lemma directly, along the lines of Gale [141].

Suppose that $Ax = b$ has no non-negative solution. If, indeed, it has no solution whatsoever, then $b = Ax + w$, where $w \neq 0$ and $A^T w = 0$. Then we take $y = -w/||w||_2^2$. So suppose that $Ax = b$ does have solutions, but not any non-negative ones. The approach is to use induction on the number of columns of the matrix involved in the lemma.

If A has only one column, denoted a^1, then $Ax = b$ can be written as

$$x_1 a^1 = b.$$

Assuming that there are no non-negative solutions, it must follow that $x_1 < 0$. We take $y = -b$. Then

$$b^T y = -b^T b = -||b||_2^2 < 0,$$

while

$$A^T y = (a^1)^T (-b) = \frac{-1}{x_1} b^T b > 0.$$

Now assume that the lemma holds whenever the involved matrix has no more than $m - 1$ columns. We show the same is true for m columns.

If there is no non-negative solution of the system $Ax = b$, then clearly there are no non-negative real numbers $x_1, x_2, ..., x_{m-1}$ such that

$$x_1 a^1 + x_2 a^2 + ... + x_{m-1} a^{m-1} = b,$$

where a^j denotes the jth column of the matrix A. By the induction hypothesis, there must be a vector v with

$$(a^j)^T v \geq 0,$$

for $j = 1, ..., m-1$, and $b^T v < 0$. If it happens that $(a^m)^T v \geq 0$ also, then we are done. If, on the other hand, we have $(a^m)^T v < 0$, then let

$$c^j = (a^j)^T a^m - (a^m)^T a^j,\ j = 1, ..., m-1,$$

and

$$d = (b^T v) a^m - ((a^m)^T v) b.$$

Then there are no non-negative real numbers $z_1, ..., z_{m-1}$ such that

$$z_1 c^1 + z_2 c^2 + ... + z_{m-1} c^{m-1} = d, \tag{9.1}$$

since, otherwise, it would follow from simple calculations that

$$\frac{-1}{(a^m)^T v}\Big([\sum_{j=1}^{m-1} z_j((a^j)^T v)] - b^T v\Big) a^m - \sum_{j=1}^{m-1} z_j((a^m)^T v) a^j = b.$$

Close inspection of this shows all the coefficients to be non-negative, which implies that the system $Ax = b$ has a non-negative solution, contrary to our assumption. It follows, therefore, that there can be no non-negative solution to the system in Equation (9.1).

By the induction hypothesis, it follows that there is a vector u such that

$$(c^j)^T u \geq 0,\ j = 1, ..., m-1,$$

and

$$d^T u < 0.$$

Now let

$$y = ((a^m)^T u) v - ((a^m)^T v) u.$$

We can easily verify that

$$(a^j)^T y = (c^j)^T u \geq 0,\ j = 1, ..., m-1,$$

$$b^T y = d^T u < 0,$$

and

$$(a^m)^T y = 0,$$

so that

$$A^T y \geq 0,$$

and

$$b^T y < 0.$$

This completes the proof.

9.3 Linear Programming

We begin with an example.

9.3.1 An Example

Consider the problem of maximizing the function $f(x_1, x_2) = x_1 + 2x_2$, over all $x_1 \geq 0$ and $x_2 \geq 0$, for which the inequalities

$$x_1 + x_2 \leq 40,$$

and

$$2x_1 + x_2 \leq 60$$

are satisfied. The set of points satisfying all four inequalities is the quadrilateral with vertices $(0, 0)$, $(30, 0)$, $(20, 20)$, and $(0, 40)$; draw a picture. Since the level curves of the function f are straight lines, the maximum value must occur at one of these vertices; in fact, it occurs at $(0, 40)$ and the maximum value of f over the constraint set is 80. Rewriting the problem as minimizing the function $-x_1 - 2x_2$, subject to $x_1 \geq 0$, $x_2 \geq 0$,

$$-x_1 - x_2 \geq -40,$$

and

$$-2x_1 - x_2 \geq -60,$$

the problem is now in what is called *primal canonical form.*

9.3.2 Canonical and Standard Forms

Let b and c be fixed vectors and A a fixed matrix. The problem

$$\text{minimize } z = c^T x, \text{ subject to } Ax \geq b, x \geq 0 \text{ (PC)} \tag{9.2}$$

is the so-called *primary problem* of LP, in *canonical form.* The *dual problem* in canonical form is

$$\text{maximize } w = b^T y, \text{ subject to } A^T y \leq c, y \geq 0. \text{ (DC)} \tag{9.3}$$

The primary problem, in *standard form*, is

$$\text{minimize } z = c^T x, \text{ subject to } Ax = b, x \geq 0 \text{ (PS)} \tag{9.4}$$

with the dual problem in standard form given by

$$\text{maximize } w = b^T y, \text{ subject to } A^T y \leq c. \text{ (DS)} \tag{9.5}$$

Notice that the dual problem in standard form does not require that y be nonnegative. Note also that (PS) makes sense only if the system $Ax = b$ has solutions. For that reason, we shall assume, for the standard problems, that the I by J matrix A has at least as many columns as rows, so $J \geq I$, and A has full rank I.

If we are given the primary problem in canonical form, we can convert it to standard form by augmenting the variables, that is, by defining

$$u_i = (Ax)_i - b_i, \tag{9.6}$$

for $i = 1, ..., I$, and rewriting $Ax \geq b$ as

$$\tilde{A}\tilde{x} = b, \tag{9.7}$$

for $\tilde{A} = \begin{bmatrix} A & -I \end{bmatrix}$ and $\tilde{x} = [x^T u^T]^T$.

If we are given the primary problem in standard form, we can convert it to canonical form by writing the equations as inequalities, that is, by replacing $Ax = b$ with the two matrix inequalities $Ax \geq b$, and $(-A)x \geq -b$.

9.3.3 Weak Duality

Consider the problems (PS) and (DS). Say that x is *feasible* if $x \geq 0$ and $Ax = b$. Let F be the set of feasible x. Say that y is *feasible* if $A^T y \leq c$. The *Weak Duality Theorem* is the following:

Theorem 9.10 *Let x and y be feasible vectors. Then*

$$z = c^T x \geq b^T y = w. \tag{9.8}$$

Corollary 9.1 *If z is not bounded below, then there are no feasible y.*

Corollary 9.2 *If x and y are both feasible, and $z = w$, then both x and y are optimal for their respective problems.*

The proof of the theorem and its corollaries are left as exercises.

The nonnegative quantity $c^T x - b^T y$ is called the *duality gap*. The *complementary slackness condition* says that, for optimal x and y, we have

$$x_j(c_j - (A^T y)_j) = 0, \tag{9.9}$$

for each j, which says that the duality gap is zero. Primal-dual algorithms for solving linear programming problems are based on finding sequences $\{x^k\}$ and $\{y^k\}$ that drive the duality gap down to zero [209].

9.3.4 Strong Duality

The *Strong Duality Theorems* make a stronger statement. The following theorems are well known examples.

Theorem 9.11 *If one of the problems (PS) or (DS) has an optimal solution, then so does the other and $z = w$ for the optimal vectors.*

Theorem 9.12 Gale's Strong Duality Theorem*[141] If both problems (PC) and (DC) have feasible solutions, then both have optimal solutions and the optimal values are equal.*

Proof: We show that there are non-negative vectors x and y such that $Ax \geq b$, $A^T y \leq c$, and $b^T y - c^T x \geq 0$. It will then follow that $z = c^T x = b^T y = w$, so that x and y are both optimal. In matrix notation, we want to find $x \geq 0$ and $y \geq 0$ such that

$$\begin{bmatrix} A & 0 \\ 0 & -A^T \\ -c^T & b^T \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \geq \begin{bmatrix} b \\ -c \\ 0 \end{bmatrix}. \tag{9.10}$$

We assume that there are no $x \geq 0$ and $y \geq 0$ for which the inequalities in (9.10) hold. Then, according to Theorem 9.7, there are non-negative vectors s and t, and non-negative scalar ρ such that

$$\begin{bmatrix} -A^T & 0 & c \\ 0 & A & -b \end{bmatrix} \begin{bmatrix} s \\ t \\ \rho \end{bmatrix} \geq 0, \tag{9.11}$$

and

$$\begin{bmatrix} -b^T & c^T & 0 \end{bmatrix} \begin{bmatrix} s \\ t \\ \rho \end{bmatrix} < 0. \tag{9.12}$$

Note that ρ cannot be zero, for then we would have $A^T s \leq 0$ and $At \geq 0$. Taking feasible vectors x and y, we would find that $s^T Ax \leq 0$, which implies that $b^T s \leq 0$, and $t^T A^T y \geq 0$, which implies that $c^T t \geq 0$. Therefore, we could not also have $c^T t - b^T s < 0$.

Writing out the inequalities, we have

$$\rho c^T t \geq s^T At \geq s^T(\rho b) = \rho s^T b.$$

Using $\rho > 0$, we find that

$$c^T t \geq b^T s,$$

which is a contradiction. Therefore, there do exist $x \geq 0$ and $y \geq 0$ such that $Ax \geq b$, $A^T y \leq c$, and $b^T y - c^T x \geq 0$. ∎

In his book [141] Gale uses his strong duality theorem to obtain a proof of the *min-max* theorem in game theory (see [70]).

Part III

Algorithms

Chapter 10

Fixed-Point Methods

10.1 Chapter Summary

In a broad sense, all iterative algorithms generate a sequence $\{x^k\}$ of vectors. The sequence may converge for any starting vector x^0, or may converge only if the x^0 is sufficiently close to a solution. The limit, when it exists, may depend on x^0, and may, or may not, solve the original problem. Convergence to the limit may be slow and the algorithm may need to be accelerated. The algorithm may involve measured data. The limit may be sensitive to noise in the data and the algorithm may need to be regularized to lessen this sensitivity. The algorithm may be quite general, applying to all problems in a broad class, or it may be tailored to the problem at hand. Each step of the algorithm may be costly, but only a few steps generally needed to produce a suitable approximate answer, or, each step may be easily performed, but many such steps needed. Although convergence of an algorithm is important, theoretically, sometimes in practice only a few iterative steps are used. In this chapter we consider several classes of operators that play important roles in applied linear algebra.

10.2 Operators

A function $T : \mathbb{R}^J \to \mathbb{R}^J$ is often called an *operator* on $\mathbb{R}^J$. For most of the iterative algorithms we shall consider, the iterative step is

$$x^{k+1} = Tx^k, \tag{10.1}$$

for some operator T. If T is a continuous operator (and it usually is), and the sequence $\{T^k x^0\}$ converges to $\hat{x}$, then $T\hat{x} = \hat{x}$, that is, $\hat{x}$ is a *fixed point* of the operator T. We denote by $\text{Fix}(T)$ the set of fixed points of T. The convergence of the iterative sequence $\{T^k x^0\}$ will depend on the properties of the operator T.

10.3 Contractions

Contraction operators are perhaps the best known class of operators associated with iterative algorithms.

10.3.1 Lipschitz Continuity

Definition 10.1 *An operator T on $\mathbb{R}^J$ is* Lipschitz continuous*, with respect to a vector norm $||\cdot||$, or L-Lipschitz, if there is a positive constant L such that*

$$||Tx - Ty|| \leq L||x - y||, \tag{10.2}$$

for all x and y in $\mathbb{R}^J$.

10.3.1.1 An Example: Bounded Derivative

We know from the Mean Value Theorem that, for any differentiable function $f : \mathbb{R} \to \mathbb{R}$,

$$f(x) - f(y) = f'(c)(x - y),$$

where c is between x and y. Suppose that there is a positive constant L such that $|f'(c)| \leq L$, for all real c. Then

$$|f(x) - f(y)| \leq L|x - y|,$$

for all x and y. Therefore, the function f is L-Lipschitz. The function $f(x) = \frac{1}{2}\cos(x)$ is $\frac{1}{2}$-Lipschitz.

More generally, if f is a real-valued differentiable function of J real variables, that is, $f : \mathbb{R}^J \to \mathbb{R}$, and the gradient satisfies $\|\nabla f(c)\|_2 \leq L$ for all c in $\mathbb{R}^J$, then

$$|f(x) - f(y)| \leq L\|x - y\|_2,$$

so that f is L-Lipschitz, with respect to the 2-norm.

10.3.1.2 Another Example: Lipschitz Gradients

If $f : \mathbb{R}^J \to \mathbb{R}$ is twice differentiable and $\|\nabla^2 f(c)\|_2 \leq L$, for all x, then $T = \nabla f$ is L-Lipschitz, with respect to the 2-norm.

10.3.2 Non-expansive Operators

Definition 10.2 *If, for some norm, T is L-Lipschitz for $L = 1$, then T is said to be* non-expansive *(ne), with respect to the given norm; in other words, $\|Tx - Ty\| \leq \|x - y\|$, for all x and y.*

Lemma 10.1 *Let $T : \mathbb{R}^J \to \mathbb{R}^J$ be a non-expansive operator, with respect to the 2-norm. Then the set F of fixed points of T is a convex set.*

Proof: If the set of fixed points of T is empty, or contains just a single point, then it is trivially convex, since there are no two points in it for which the convexity condition fails. So assume that the set of fixed points contains at least two members. Select two distinct points a and b in F, a scalar α in the open interval $(0, 1)$, and let $c = \alpha a + (1-\alpha)b$. We show that $Tc = c$. Note that

$$a - c = \frac{1-\alpha}{\alpha}(c - b).$$

We have

$$\|a-b\|_2 = \|a-Tc+Tc-b\|_2 \leq \|a-Tc\|_2+\|Tc-b\|_2 = \|Ta-Tc\|_2+\|Tc-Tb\|_2$$

$$\leq \|a - c\|_2 + \|c - b\|_2 = \|a - b\|_2;$$

the last equality follows since $a - c$ is a multiple of $(c - b)$. From this, we conclude that

$$\|a - Tc\|_2 = \|a - c\|_2,$$

$$\|Tc - b\|_2 = \|c - b\|_2,$$

and that $a - Tc$ and $Tc - b$ are positive multiples of one another, that is, there is $\beta > 0$ such that

$$a - Tc = \beta(Tc - b),$$

or

$$Tc = \frac{1}{1+\beta}a + \frac{\beta}{1+\beta}b = \gamma a + (1-\gamma)b.$$

Then inserting $c = \alpha a + (1 - \alpha)b$ and $Tc = \gamma a + (1 - \gamma)b$ into

$$\|Tc - b\|_2 = \|c - b\|_2,$$

we find that $\gamma = \alpha$ and so $Tc = c$. ∎

We want to find properties of an operator T that guarantee that the sequence of iterates $\{T^k x_0\}$ will converge to a fixed point of T, for any x^0, whenever fixed points exist. Being non-expansive is not enough; the non-expansive operator $T = -I$, where $Ix = x$ is the identity operator, has the fixed point $x = 0$, but the sequence $\{T^k x^0\}$ converges only if $x^0 = 0$.

10.3.3 Strict Contractions

One property that guarantees not only that the iterates converge, but that there is a fixed point is the property of being a strict contraction.

Definition 10.3 *An operator T on $\mathbb{R}^J$ is a* strict contraction *(sc), with respect to a vector norm $||\cdot||$, if there is $r \in (0,1)$ such that*

$$||Tx - Ty|| \leq r||x - y||, \tag{10.3}$$

for all vectors x and y.

For example, if the operator T is L-Lipschitz for some $L < 1$, then T is a strict contraction. Therefore, if $f : \mathbb{R}^J \to \mathbb{R}$ is differentiable and ∇f is L-Lipschitz for some $L < 1$, for all x, then $T = \nabla f$ is a strict contraction.

For strict contractions, we have the Banach-Picard Theorem [122]:

Theorem 10.1 *Let T be sc. Then, there is a unique fixed point of T and, for any starting vector x^0, the sequence $\{T^k x^0\}$ converges to the fixed point.*

The key step in the proof is to show that $\{x^k\}$ is a Cauchy sequence, therefore, it has a limit.

Corollary 10.1 *If T^n is a strict contraction, for some positive integer n, then T has a fixed point.*

Proof: Suppose that $T^n\hat{x} = \hat{x}$. Then

$$T^n T\hat{x} = TT^n\hat{x} = T\hat{x},$$

so that both $\hat{x}$ and $T\hat{x}$ are fixed points of T^n. But T^n has a unique fixed point. Therefore, $T\hat{x} = \hat{x}$. ∎

10.3.4 Eventual Strict Contractions

Consider the problem of finding x such that $x = e^{-x}$. We can see from the graphs of $y = x$ and $y = e^{-x}$ that there is a unique solution, which we shall denote by z. It turns out that $z = 0.56714329040978....$ Let us try to find z using the iterative sequence $x_{k+1} = e^{-x_k}$, starting with some real x_0. Note that we always have $x_k > 0$ for $k = 1, 2, ...$, even if $x_0 < 0$. The operator here is $Tx = e^{-x}$, which, for simplicity, we view as an operator on the non-negative real numbers.

Since the derivative of the function $f(x) = e^{-x}$ is $f'(x) = -e^{-x}$, we have $|f'(x)| \leq 1$, for all non-negative x, so T is non-expansive. But we do not have $|f'(x)| \leq r < 1$, for all non-negative x; therefore, T is a not a strict contraction, when considered as an operator on the non-negative real numbers.

If we choose $x_0 = 0$, then $x_1 = 1$, $x_2 = 0.368$, approximately, and so on. Continuing this iteration a few more times, we find that after about $k = 14$, the value of x_k settles down to 0.567, which is the answer, to three decimal places. The same thing is seen to happen for any positive starting points x_0.

It would seem that T has another property, besides being non-expansive, that is forcing convergence. What is it?

From the fact that $1 - e^{-x} \leq x$, for all real x, with equality if and only if $x = 0$, we can show easily that, for $r = \max\{e^{-x_1}, e^{-x_2}\}$,

$$|z - x_{k+1}| \leq r|z - x_k|,$$

for $k = 3, 4, ...$ Since $r < 1$, it follows, just as in the proof of the Banach-Picard Theorem, that $\{x_k\}$ is a Cauchy sequence and therefore converges. The limit must be a fixed point of T, so the limit must be z.

Although the operator T is not a strict contraction, with respect to the non-negative numbers, once we begin to calculate the sequence of iterates the operator T effectively becomes a strict contraction, with respect to the vectors of the particular sequence being constructed, and so the sequence converges to a fixed point of T. We cannot conclude from this that T has a unique fixed point, as we can in the case of a strict contraction; we must decide that by other means.

Note that the operator $Tx = e^{-x}$ is paracontractive, a notion that we shall discuss later in this chapter, so convergence to the fixed point is also a consequence of the EKN Theorem 10.6, also to be discussed later.

Ex. 10.1 *Show that a strict contraction can have at most one fixed point.*

Ex. 10.2 *Let T be sc. Show that the sequence $\{T^k x_0\}$ is a Cauchy sequence. Hint: consider*

$$||x^k - x^{k+n}|| \leq ||x^k - x^{k+1}|| + ... + ||x^{k+n-1} - x^{k+n}||, \tag{10.4}$$

and use

$$||x^{k+m} - x^{k+m+1}|| \leq r^m ||x^k - x^{k+1}||. \tag{10.5}$$

Since $\{x^k\}$ is a Cauchy sequence, it has a limit, say $\hat{x}$. Let $e^k = \hat{x} - x^k$. Show that $\{e^k\} \to 0$, as $k \to +\infty$, so that $\{x^k\} \to \hat{x}$. Finally, show that $T\hat{x} = \hat{x}$.

Ex. 10.3 *Suppose that we want to solve the equation*

$$x = \frac{1}{2}e^{-x}.$$

Let $Tx = \frac{1}{2}e^{-x}$ for x in $\mathbb{R}$. Show that T is a strict contraction, when restricted to non-negative values of x, so that, provided we begin with $x^0 > 0$, the sequence $\{x^k = Tx^{k-1}\}$ converges to the unique solution of the equation. Hint: use the mean value theorem from calculus.

10.3.5 Instability

Suppose we rewrite the equation $e^{-x} = x$ as $x = -\log x$, and define $Tx = -\log x$, for $x > 0$. Now our iterative scheme becomes $x_{k+1} = Tx_k = -\log x_k$. A few calculations will convince us that the sequence $\{x_k\}$ is diverging away from the correct answer, not converging to it. The lesson here is that we cannot casually reformulate our problem as a fixed-point problem and expect the iterates to converge to the answer. What matters is the behavior of the operator T.

10.4 Gradient Descent

We provide an elementary proof of the following theorem.

Theorem 10.2 *Let $f : \mathbb{R}^J \to \mathbb{R}$ be convex and differentiable, with ∇f L-Lipschitz. For $0 < \gamma < \frac{1}{L}$, let $T = I - \gamma\nabla f$. If T has fixed points, then the sequence $\{x^k\}$ given by $x^k = Tx^{k-1}$ converges to a fixed point of T.*

The iterative step is given by

$$x^k = x^{k-1} - \gamma\nabla f(x^{k-1}). \tag{10.6}$$

It is a consequence of the Krasnoselskii-Mann Theorem for averaged operators that convergence holds for $0 < \gamma < \frac{2}{L}$. The proof given here employs sequential unconstrained minimization and avoids using the non-trivial theorem that tells us that, because the operator $\frac{1}{L}\nabla f$ is non-expansive, it is firmly non-expansive.

10.4.1 Using Sequential Unconstrained Minimization

For each $k = 1, 2, ...$ let

$$G_k(x) = f(x) + \frac{1}{2\gamma}\|x - x^{k-1}\|_2^2 - D_f(x, x^{k-1}), \tag{10.7}$$

where

$$D_f(x, x^{k-1}) = f(x) - f(x^{k-1}) - \langle\nabla f(x^{k-1}), x - x^{k-1}\rangle. \tag{10.8}$$

Since $f(x)$ is convex, $D_f(x, y) \geq 0$ for all x and y and is the Bregman distance formed from the function f [29]. The x^k that minimizes $G_k(x)$ is given by Equation (10.6).

The auxiliary function

$$g_k(x) = \frac{1}{2\gamma}\|x - x^{k-1}\|_2^2 - D_f(x, x^{k-1}) \tag{10.9}$$

can be rewritten as

$$g_k(x) = D_h(x, x^{k-1}), \tag{10.10}$$

where

$$h(x) = \frac{1}{2\gamma}\|x\|_2^2 - f(x). \tag{10.11}$$

Therefore, $g_k(x) \geq 0$ whenever $h(x)$ is a convex function.

We know that $h(x)$ is convex if and only if

$$\langle \nabla h(x) - \nabla h(y), x - y\rangle \geq 0, \tag{10.12}$$

for all x and y. This is equivalent to

$$\frac{1}{\gamma}\|x - y\|_2^2 - \langle \nabla f(x) - \nabla f(y), x - y\rangle \geq 0. \tag{10.13}$$

Since ∇f is L-Lipschitz, the inequality (10.13) holds whenever $0 < \gamma < \frac{1}{L}$.

It is easy to see that

$$G_k(x) - G_k(x^k) = D_f(x, x^k) + D_h(x, x^k)$$

$$= \frac{1}{2\gamma}\|x - x^k\|_2^2 \geq D_h(x, x^k) = g_{k+1}(x). \tag{10.14}$$

We turn now to the behavior of the sequence $\{x^k\}$.

10.4.2 Proving Convergence

Let $Tz = z$ for the operator $T = I - \gamma\nabla f$, so that $\nabla f(z) = 0$ and z is a global minimizer of $f(x)$. We have

$$\frac{1}{2\gamma}\|z - x^k\|_2^2 = G_{k+1}(z) - G_{k+1}(x^{k+1})$$

$$= f(z) - f(x^{k+1}) + g_{k+1}(z) - g_{k+1}(x^{k+1})$$

$$\leq f(z) - f(x^{k+1}) + G_k(z) - G_k(x^k) - g_{k+1}(x^{k+1}), \tag{10.15}$$

so that

$$\Big(G_k(z) - G_k(x^k)\Big) - \Big(G_{k+1}(z) - G_{k+1}(x^{k+1})\Big) = \frac{1}{2\gamma}\Big(\|z - x^{k-1}\|_2^2 - \|z - x^k\|_2^2\Big)$$

$$\geq f(x^{k+1}) - f(z) + g_{k+1}(x^{k+1}) \geq 0. \tag{10.16}$$

It follows that the sequence $\{\|z - x^k\|_2^2\}$ is decreasing, and the sequences $\{f(x^{k+1}) - f(z)\}$ and $\{g_{k+1}(x^{k+1})\}$ converge to zero. The sequence $\{x^k\}$ is therefore bounded and has a subsequence converging to x^*, with $f(x^*) = f(z)$. We replace the generic fixed point z with x^* in the expressions above, and conclude that the sequence $\{\|x^* - x^k\|_2^2\}$ converges to zero. This completes the proof of the theorem.

10.4.3 An Example: Least Squares

We consider the problem of minimizing the function $f(x) = \frac{1}{2}\|Ax - b\|_2^2$ to get a least squares solution of the linear system $Ax = b$. The gradient of $f(x)$ is

$$\nabla f(x) = A^T Ax - A^T b, \tag{10.17}$$

and is L-Lipschitz for $L = \rho(A^T A)$, the largest eigenvalue of the matrix $A^T A$.

The Landweber iterative algorithm is defined by the iterative step

$$x^k = x^{k-1} - \gamma A^T (Ax^{k-1} - b). \tag{10.18}$$

It follows from our previous discussion that the sequence $\{x^k\}$ converges to a least squares solution for any γ satisfying the inequalities $0 < \gamma < \frac{1}{\rho(A^T A)}$.

10.5 Two Useful Identities

The notions of non-expansive operator and strict contraction make sense for any norm. In what follows, when we speak of orthogonal projections, averaged operators or firmly non-expansive operators, it will be understood that we are speaking about the 2-norm, since the definitions of such operators involve the inner product.

The identities in the next two lemmas apply to the 2-norm and relate an arbitrary operator T to its complement, $G = I - T$, where I denotes the identity operator. These identities will allow us to transform properties of T into properties of G that may be easier to work with. A simple calculation is all that is needed to establish the following lemma.

Lemma 10.2 *Let T be an arbitrary operator T on $\mathbb{R}^J$ and $G = I - T$. Then*

$$||x - y||_2^2 - ||Tx - Ty||_2^2 = 2(\langle Gx - Gy, x - y\rangle) - ||Gx - Gy||_2^2. \tag{10.19}$$

Proof: We use

$$\|a+b\|_2^2 = \langle a+b, a+b\rangle = \langle a,a\rangle + 2\langle a,b\rangle + \langle b,b\rangle = \|a\|_2^2 + 2\langle a,b\rangle + \|b\|_2^2.$$

We write

$$\|x-y\|_2^2 = \|Tx - Ty + (x-Tx) - (y-Ty)\|_2^2$$
$$= \|Tx-Ty\|^2 + 2\langle Tx-Ty, Gx-Gy\rangle + \|Gx-Gy\|_2^2.$$

Then write

$$2\langle Tx-Ty, Gx-Gy\rangle = 2\langle x-y-Gx+Gy, Gx-Gy\rangle = \langle Gx-Gy, x-y\rangle - \|Gx-Gy\|_2^2.$$

∎

Lemma 10.3 *Let T be an arbitrary operator T on $\mathbb{R}^J$ and $G = I - T$. Then*

$$\langle Tx-Ty, x-y\rangle - ||Tx-Ty||_2^2 =$$

$$\langle Gx-Gy, x-y\rangle - ||Gx-Gy||_2^2. \tag{10.20}$$

Proof: Use the previous lemma. ∎

10.6 Orthogonal Projection Operators

If C is a closed, non-empty convex set in $\mathbb{R}^J$, and x is any vector, then, as we have seen, there is a unique point $P_C x$ in C closest to x, in the sense of the Euclidean distance. This point is called the orthogonal projection of x onto C. If C is a subspace, then we can get an explicit description of $P_C x$ in terms of x; for general convex sets C, however, we will not be able to express $P_C x$ explicitly, and certain approximations will be needed. Orthogonal projection operators are central to our discussion, and, in this overview, we focus on problems involving convex sets, algorithms involving orthogonal projection onto convex sets, and classes of operators derived from properties of orthogonal projection operators.

10.6.1 Properties of the Operator P_C

Although we usually do not have an explicit expression for $P_C x$, we can, however, characterize $P_C x$ as the unique member of C for which

$$\langle P_C x - x, c - P_C x\rangle \geq 0, \tag{10.21}$$

for all c in C; see Proposition 8.4.

10.6.1.1 P_C is Non-expansive

Recall that an operator T is non-expansive (ne), with respect to a given norm, if, for all x and y, we have

$$||Tx - Ty|| \leq ||x - y||. \tag{10.22}$$

Lemma 10.4 *The orthogonal projection operator $T = P_C$ is non-expansive, with respect to the Euclidean norm, that is,*

$$||P_C x - P_C y||_2 \leq ||x - y||_2, \tag{10.23}$$

for all x and y.

Proof: Use Inequality (10.21) to get

$$\langle P_C y - P_C x, P_C x - x \rangle \geq 0, \tag{10.24}$$

and

$$\langle P_C x - P_C y, P_C y - y \rangle \geq 0. \tag{10.25}$$

Add the two inequalities to obtain

$$\langle P_C x - P_C y, x - y \rangle \geq ||P_C x - P_C y||_2^2, \tag{10.26}$$

and use the Cauchy Inequality. ∎

Because the operator P_C has multiple fixed points, P_C cannot be a strict contraction, unless the set C is a singleton set.

10.6.1.2 P_C is Firmly Non-expansive

Definition 10.4 *An operator T is said to be* firmly non-expansive *(fne) if*

$$\langle Tx - Ty, x - y \rangle \geq ||Tx - Ty||_2^2, \tag{10.27}$$

for all x and y in $\mathbb{R}^J$.

Lemma 10.5 *An operator T is fne if and only if $G = I - T$ is fne.*

Proof: Use the identity in Equation (10.20). ∎

From Equation (10.26), we see that the operator $T = P_C$ is not simply ne, but fne, as well. A good source for more material on these topics is the book by Goebel and Reich [147].

Proposition 10.1 *An operator F is firmly non-expansive if and only if $F = \frac{1}{2}(I + N)$, for some non-expansive operator N.*

Ex. 10.4 *Prove Proposition 10.1.*

10.6.1.3 The Search for Other Properties of P_C

The class of non-expansive operators is too large for our purposes; the operator $Tx = -x$ is non-expansive, but the sequence $\{T^k x^0\}$ does not converge, in general, even though a fixed point, $x = 0$, exists. The class of firmly non-expansive operators is too small for our purposes. Although the convergence of the iterative sequence $\{T^k x^0\}$ to a fixed point does hold for firmly non-expansive T, whenever fixed points exist, the product of two or more fne operators need not be fne; that is, the class of fne operators is not *closed to finite products*. This poses a problem, since, as we shall see, products of orthogonal projection operators arise in several of the algorithms we wish to consider. We need a class of operators smaller than the ne ones, but larger than the fne ones, closed to finite products, and for which the sequence of iterates $\{T^k x^0\}$ will converge, for any x^0, whenever fixed points exist. The class we shall consider is the class of *averaged* operators.

For the remainder of this chapter the term *non-expansive* will mean with respect to the Euclidean norm, unless otherwise indicated from the context.

10.7 Averaged Operators

The term 'averaged operator' appears in the work of Baillon, Bruck and Reich [34, 9]. There are several ways to define averaged operators. One way is in terms of the complement operator.

Definition 10.5 *An operator G on $\mathbb{R}^J$ is called* ν-inverse strongly monotone *(ν-ism)[148] (also called co-coercive in [98]) if there is $\nu > 0$ such that*

$$\langle Gx - Gy, x - y \rangle \geq \nu ||Gx - Gy||_2^2. \tag{10.28}$$

Lemma 10.6 *An operator T is ne if and only if its complement $G = I - T$ is $\frac{1}{2}$-ism, and T is fne if and only if G is 1-ism, and if and only if G is fne. Also, T is ne if and only if $F = (I + T)/2$ is fne. If G is ν-ism and $\gamma > 0$ then the operator γG is $\frac{\nu}{\gamma}$-ism.*

Ex. 10.5 *Prove Lemma 10.6.*

Definition 10.6 *An operator T is called* averaged *(av) if $G = I - T$ is ν-ism for some $\nu > \frac{1}{2}$. If G is $\frac{1}{2\alpha}$-ism, for some $\alpha \in (0, 1)$, then we say that T is α-av.*

It follows that every av operator is ne, with respect to the Euclidean norm, and every fne operator is av.

The averaged operators are sometimes defined in a different, but equivalent, way, using the following characterization of av operators.

Lemma 10.7 *An operator T is av if and only if, for some operator N that is non-expansive in the Euclidean norm, and $\alpha \in (0,1)$, we have*

$$T = (1-\alpha)I + \alpha N.$$

Consequently, the operator T is av if and only if, for some α in $(0,1)$, the operator

$$N = \frac{1}{\alpha}T - \frac{1-\alpha}{\alpha}I = I - \frac{1}{\alpha}(I-T) = I - \frac{1}{\alpha}G$$

is non-expansive.

Proof: We assume first that there is $\alpha \in (0,1)$ and ne operator N such that $T = (1-\alpha)I + \alpha N$, and so $G = I - T = \alpha(I-N)$. Since N is ne, $I-N$ is $\frac{1}{2}$-ism and $G = \alpha(I-N)$ is $\frac{1}{2\alpha}$-ism. Conversely, assume that G is ν-ism for some $\nu > \frac{1}{2}$. Let $\alpha = \frac{1}{2\nu}$ and write $T = (1-\alpha)I + \alpha N$ for $N = I - \frac{1}{\alpha}G$. Since $I - N = \frac{1}{\alpha}G$, $I-N$ is $\alpha\nu$-ism. Consequently $I-N$ is $\frac{1}{2}$-ism and N is ne. ∎

An averaged operator is easily constructed from a given ne operator N by taking a convex combination of N and the identity I. The beauty of the class of av operators is that it contains many operators, such as P_C, that are not originally defined in this way. As we shall see shortly, finite products of averaged operators are again averaged, so the product of finitely many orthogonal projections is av.

We present now the fundamental properties of averaged operators, in preparation for the proof that the class of averaged operators is closed to finite products.

Note that we can establish that a given operator A is av by showing that there is an α in the interval $(0,1)$ such that the operator

$$\frac{1}{\alpha}(A - (1-\alpha)I) \tag{10.29}$$

is ne. Using this approach, we can easily show that if T is sc, then T is av.

Lemma 10.8 *Let $T = (1-\alpha)A + \alpha N$ for some $\alpha \in (0,1)$. If A is averaged and N is non-expansive then T is averaged.*

Proof: Let $A = (1-\beta)I + \beta M$ for some $\beta \in (0,1)$ and ne operator M. Let $1-\gamma = (1-\alpha)(1-\beta)$. Then we have

$$T = (1-\gamma)I + \gamma[(1-\alpha)\beta\gamma^{-1}M + \alpha\gamma^{-1}N]. \tag{10.30}$$

Since

$$(1-\alpha)\beta\gamma^{-1} + \alpha\gamma^{-1} = 1,$$

and both M and N are ne, the operator

$$K = (1-\alpha)\beta\gamma^{-1}M + \alpha\gamma^{-1}N,$$

which is the convex combination of two ne operators, is again ne, so that T is averaged. ∎

Corollary 10.2 *If A and B are av and α is in the interval $[0,1]$, then the operator $T = (1-\alpha)A + \alpha B$ formed by taking the convex combination of A and B is av.*

Corollary 10.3 *Let $T = (1-\alpha)F + \alpha N$ for some $\alpha \in (0,1)$. If F is fne and N is ne then T is averaged.*

Ex. 10.6 *Show that, if the operator T is α-av and $1 > \beta > \alpha$, then T is β-av.*

The orthogonal projection operators P_H onto hyperplanes $H = H(a,\gamma)$ are sometimes used with *relaxation*, which means that P_H is replaced by the operator

$$T = (1-\omega)I + \omega P_H, \tag{10.31}$$

for some ω in the interval $(0,2)$. Clearly, if ω is in the interval $(0,1)$, then T is av, by definition, since P_H is ne. We want to show that, even for ω in the interval $[1,2)$, T is av. To do this, we consider the operator $R_H = 2P_H - I$, which is reflection through H; that is,

$$P_H x = \frac{1}{2}(x + R_H x), \tag{10.32}$$

for each x.

Lemma 10.9 *The operator $R_H = 2P_H - I$ is an isometry; that is,*

$$||R_H x - R_H y||_2 = ||x - y||_2, \tag{10.33}$$

for all x and y, so that R_H is ne.

Lemma 10.10 *For $\omega = 1 + \gamma$ in the interval $[1,2)$, we have*

$$(1-\omega)I + \omega P_H = \alpha I + (1-\alpha)R_H, \tag{10.34}$$

for $\alpha = \frac{1-\gamma}{2}$; therefore, $T = (1-\omega)I + \omega P_H$ is av.

The product of finitely many ne operators is again ne, while the product of finitely many fne operators, even orthogonal projections, need not be fne. It is a helpful fact that the product of finitely many av operators is again av.

If $A = (1-\alpha)I + \alpha N$ is averaged and B is averaged then $T = AB$ has the form $T = (1-\alpha)B + \alpha NB$. Since B is av and NB is ne, it follows from Lemma 10.8 that T is averaged. Summarizing, we have

Proposition 10.2 *If A and B are averaged, then $T = AB$ is averaged.*

10.7.1 Gradient Operators

Another type of operator that is averaged can be derived from gradient operators.

Definition 10.7 *An operator T on $\mathbb{R}^J$ is* monotone *if*

$$\langle Tx - Ty, x - y \rangle \geq 0, \tag{10.35}$$

for all x and y.

Firmly non-expansive operators on $\mathbb{R}^J$ are monotone operators. Let $g(x) : \mathbb{R}^J \to \mathbb{R}$ be a differentiable convex function and $f(x) = \nabla g(x)$ its gradient. The operator ∇g is also monotone. If ∇g is non-expansive, then it can be shown that ∇g is fne (see the chapter on Convex Functions in [70]). If, for some $L > 0$, ∇g is L-Lipschitz, for the 2-norm, that is,

$$||\nabla g(x) - \nabla g(y)||_2 \leq L||x - y||_2, \tag{10.36}$$

for all x and y, then $\frac{1}{L}\nabla g$ is ne, therefore fne, and the operator $T = I - \gamma \nabla g$ is av, for $0 < \gamma < \frac{2}{L}$.

10.7.2 The Krasnoselskii-Mann Theorem

For any operator T that is averaged, convergence of the sequence $\{T^k x^0\}$ to a fixed point of T, whenever fixed points of T exist, is guaranteed by the Krasnoselskii-Mann (KM) Theorem [201]:

Theorem 10.3 *Let T be averaged. Then the sequence $\{T^k x^0\}$ converges to a fixed point of T, whenever Fix(T) is non-empty.*

Proof: Let z be a fixed point of non-expansive operator N and let $\alpha \in (0, 1)$. Let $T = (1-\alpha)I + \alpha N$, so the iterative step becomes

$$x^{k+1} = Tx^k = (1-\alpha)x^k + \alpha N x^k. \tag{10.37}$$

The identity in Equation (10.19) is the key to proving Theorem 10.3.

Using $Tz = z$ and $(I - T)z = 0$ and setting $G = I - T$ we have

$$||z - x^k||_2^2 - ||Tz - x^{k+1}||_2^2 = 2\langle Gz - Gx^k, z - x^k\rangle - ||Gz - Gx^k||_2^2. \tag{10.38}$$

Since, by Lemma 10.7, G is $\frac{1}{2\alpha}$-ism, we have

$$||z - x^k||_2^2 - ||z - x^{k+1}||_2^2 \geq (\frac{1}{\alpha} - 1)||x^k - x^{k+1}||_2^2. \tag{10.39}$$

Consequently the sequence $\{x^k\}$ is bounded, the sequence $\{||z - x^k||_2\}$ is decreasing and the sequence $\{||x^k - x^{k+1}||_2\}$ converges to zero. Let x^* be a cluster point of $\{x^k\}$. Then we have $Tx^* = x^*$, so we may use x^* in place of the arbitrary fixed point z. It follows then that the sequence $\{||x^* - x^k||_2\}$ is decreasing; since a subsequence converges to zero, the entire sequence converges to zero. The proof is complete. ∎

A version of the KM Theorem 10.3, with variable coefficients, appears in Reich's paper [228].

10.8 Affine Linear Operators

It may not always be easy to decide if a given operator is averaged. The class of affine linear operators provides an interesting illustration of the problem.

The affine operator $Tx = Bx + d$ on $\mathbb{C}^J$ will be ne, sc, fne, or av precisely when the linear operator given by multiplication by the matrix B is the same.

10.8.1 The Hermitian Case

When B is Hermitian, we can determine if B belongs to these classes by examining its eigenvalues λ.

Theorem 10.4 *Suppose that B is an Hermitian matrix, viewed as a linear operator on $\mathbb{C}^J$. Then*

- *B is non-expansive if and only if $-1 \leq \lambda \leq 1$, for all λ;*
- *B is averaged if and only if $-1 < \lambda \leq 1$, for all λ;*
- *B is a strict contraction if and only if $-1 < \lambda < 1$, for all λ;*
- *B is firmly non-expansive if and only if $0 \leq \lambda \leq 1$, for all λ.*

Proof: Since B is linear, $G = I - B$ will be ν-ism for some $\nu > 0$ if

$$\langle Gx, x\rangle \geq \nu \|Gx\|_2^2.$$

Let $\{u^1, ..., u^J\}$ be an orthonormal basis for $\mathbb{C}^J$, consisting of eigenvectors of B, with $Bu^j = \lambda_j u^j$, for $j = 1, ..., J$. In order for B to be ne it is necessary and sufficient that $|\lambda| \leq 1$ for each eigenvalue λ of B. Order the eigenvalues so that $1 \geq \lambda_1 \geq \lambda_2 \geq ... \geq \lambda_J \geq -1$.

Let x be arbitrary in $\mathbb{C}^J$. Then there are coefficients a_j so that

$$x = \sum_{j=1}^{J} a_j u^j.$$

It follows that

$$Bx = \sum_{j=1}^{J} \lambda_j a_j u^j,$$

and

$$Gx = (I - B)x = x - Bx = \sum_{j=1}^{J} (1 - \lambda_j) a_j u^j.$$

Then

$$\langle Gx, x\rangle = \sum_{j=1}^{J} (1 - \lambda_j)|a_j|^2,$$

and

$$\|Gx\|_2^2 = \sum_{j=1}^{J} (1 - \lambda_j)^2 |a_j|^2.$$

Then G is ν-ism for $\nu^{-1} = \max\{1 - \lambda_j\}$.

Let λ be an arbitrary eigenvalue of B. Since $-1 \leq \lambda \leq 1$ always holds, we have $\nu^{-1} \leq 2$, $\nu \geq \frac{1}{2}$, and B is ne. If $-1 < \lambda \leq 1$ always holds, then $\nu > \frac{1}{2}$ and B is av. If $0 \leq \lambda \leq 1$ always holds, then $\nu \geq 1$ and B is fne. If $-1 < \lambda < 1$ always holds, then

$$\|Bx\|^2 = \sum_{j=1}^{J} |\lambda_j|^2 |a_j|^2 \leq r \sum_{j=1}^{J} |a_j|^2 = r\|x\|^2,$$

for $r = \max\{|\lambda_j|^2\} < 1$. Therefore, B is sc. The converses of these assertions are easy to prove. ∎

10.8.2 Example: Landweber's Algorithm

The Landweber algorithm for solving a complex linear system of equations $Ax = b$ has the iterative step

$$x^{k+1} = x^k - \gamma A^\dagger (Ax^k - b) = (I - \gamma A^\dagger A)x^k + A^\dagger b. \tag{10.40}$$

With $B = I - \gamma A^\dagger A$ and $d = A^\dagger b$ the iteration is $x^{k+1} = Tx^k = Bx^k + d$. The matrix B is Hermitian, so is averaged if and only if every eigenvalue λ of B lies in the interval $(-1, 1]$. This means that T is averaged for γ in the interval $(0, \frac{2}{L})$, where $L = \rho(A^\dagger A)$ is the largest eigenvalue of the matrix $A^\dagger A$.

10.8.3 What if B is not Hermitian?

In Exercise 10.7 you are asked to show that, even if B is not Hermitian, if B is av, then $|\lambda| < 1$ for all eigenvalues of B that are not equal to one. The converse is not true, though; see Exercise 10.8.

Ex. 10.7 *Show that, if B is a linear av operator, then all its eigenvalues are real and $|\lambda| < 1$ for all eigenvalues λ of B that are not equal to one.*

Ex. 10.8 *Show that having $|\lambda| < 1$ for all eigenvalues λ of B that are not equal to one does not imply that the linear operator B is av. Hint: by selecting x^0 appropriately, show that the linear operator defined by the matrix*

$$B = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$$

is not av by showing that it is not even ne.

Affine linear operators T that arise, for instance, in splitting methods for solving systems of linear equations, generally have non-Hermitian linear part B. Deciding if such operators belong to these classes is more difficult. Instead, we can ask if the operator is *paracontractive*, with respect to some norm.

10.9 Paracontractive Operators

By examining the properties of the orthogonal projection operators P_C, we were led to the useful class of averaged operators. The orthogonal projections also belong to another useful class, the paracontractions.

Definition 10.8 *An operator T is called* paracontractive *(pc), with respect to a given norm, if, for every fixed point y of T, we have*

$$||Tx - y|| < ||x - y||, \tag{10.41}$$

unless $Tx = x$.

Paracontractive operators are studied by Censor and Reich in [87].

Ex. 10.9 *Show that the operator $Tx = e^{-x}$ is a paracontraction on the non-negative reals.*

Proposition 10.3 *The operators $T = P_C$ are paracontractive, with respect to the Euclidean norm.*

Proof: It follows from Lemma 10.4 and Cauchy's Inequality that

$$||P_C x - P_C y||_2 \leq ||x - y||_2,$$

with equality if and only if

$$P_C x - P_C y = \alpha(x - y),$$

for some scalar α with $|\alpha| = 1$. If we have equality then, because

$$0 \leq \langle P_C x - P_C y, x - y \rangle = \alpha ||x - y||_2^2,$$

it follows that $\alpha = 1$, and so

$$P_C x - x = P_C y - y.$$

So if y is a fixed point of P_C and equality holds, then x is also a fixed point of P_C. ∎

When we ask if a given operator T is pc, we must specify the norm. We often construct the norm specifically for the operator involved. To illustrate, we consider the case of affine operators.

10.9.1 Diagonalizable Linear Operators

Let the matrix B be diagonalizable and let the columns of U form a basis for $\mathbb{C}^J$ consisting of eigenvectors of B. Then we have $U^{-1}BU = L$, where L is the diagonal matrix having the eigenvalues of B along its diagonal.

The characteristic polynomial of any complex J by J matrix B has J complex roots, which need not all be distinct, however. In order for B to be diagonalizable, we need to know that there is a set of J linearly independent eigenvectors of B. Lemma 6.6 tells us that when all the eigenvalues of B

are distinct B is diagonalizable. When some of the eigenvalues of B are the same, this need not be the case.

We see from Lemma 6.6 that almost all square matrices B are diagonalizable. Indeed, all Hermitian B are diagonalizable. If B has real entries, but is not symmetric, then the eigenvalues of B need not be real, and the eigenvectors of B can have non-real entries. Consequently, we must consider B as a linear operator on $\mathbb{C}^J$, if we are to talk about diagonalizability. For example, consider the real matrix

$$B = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}. \tag{10.42}$$

Its eigenvalues are $\lambda = i$ and $\lambda = -i$. The corresponding eigenvectors are $(1, i)^T$ and $(1, -i)^T$. The matrix B is then diagonalizable as an operator on $\mathbb{C}^2$, but not as an operator on $\mathbb{R}^2$.

When B is not Hermitian, it is not as easy to determine if the affine operator T is sc with respect to a given norm. Instead, we often tailor the norm to the operator T. Suppose that B is a diagonalizable matrix, that is, there is a basis for $\mathbb{R}^J$ consisting of eigenvectors of B. Let $\mathcal{A} = \{u^1, ..., u^J\}$ be such a basis, and let $Bu^j = \lambda_j u^j$, for each $j = 1, ..., J$. For each x in $\mathbb{R}^J$, there are unique coefficients a_j so that

$$x = \sum_{j=1}^{J} a_j u^j. \tag{10.43}$$

Then let

$$||x|| = \sum_{j=1}^{J} |a_j|. \tag{10.44}$$

Recall that we saw previously that once we have a basis $\mathcal{A}$ we can define an inner product, and therefore, a norm, using the vectors $[x]_{\mathcal{A}}$. That norm is the two-norm of the vectors $[x]_{\mathcal{A}}$. The norm we have just introduced is the *one-norm* of the vectors $[x]_{\mathcal{A}}$.

Lemma 10.11 *The expression $||\cdot||$ in Equation (10.44) defines a norm on $\mathbb{R}^J$. If $\rho(B) < 1$, then the affine operator T is sc, with respect to this norm.*

Ex. 10.10 *Prove Lemma 10.11.*

It is known that, for any square matrix B and any $\epsilon > 0$, there is a vector norm for which the induced matrix norm satisfies $||B|| \leq \rho(B) + \epsilon$. Therefore, if B is an arbitrary square matrix with $\rho(B) < 1$, there is a vector norm with respect to which B is sc.

We know that having $|\lambda| < 1$, unless $\lambda = 1$, does not imply that B is an av operator, so we cannot establish convergence of the sequence $\{B^k x^0\}$ by invoking the KM Theorem. However, we do have a convergence theorem for such B.

Theorem 10.5 *Let B be a diagonalizable linear operator on $\mathbb{C}^J$ whose eigenvalues satisfy the condition that $|\lambda| < 1$, unless $\lambda = 1$. Then the sequence $\{B^k x^0\}$ converges to a fixed point of B for every starting vector x^0.*

Proof: Let $\{u^1, ..., u^J\}$ be a basis for $\mathbb{C}^J$, with $Bu^j = \lambda_j u^j$, for $j = 1, ..., J$. Suppose that $\lambda_j = 1$ for $j = 1, ..., M \leq J$. Then there are coefficients a_j such that

$$x^0 = a_1 u^1 + ... + a_J u^J.$$

Then, for $k = 1, 2, ...$, we have

$$B^k x^0 = \lambda^k a_1 u^1 + ... + \lambda^k a_J u^J,$$

which converges to the fixed point $a_1 u^1 + ... + a_M u^M$, as $k \to \infty$. If there are no λ equal to one, then the sequence $\{B^k x^0\}$ converges to the zero vector. ∎

10.9.2 Linear and Affine Paracontractions

Proposition 10.4 *Let T be an affine linear operator whose linear part B is diagonalizable, and $|\lambda| < 1$ for all eigenvalues λ of B that are not equal to one. Then the operator T is pc, with respect to the norm given by Equation (10.44).*

Ex. 10.11 *Prove Proposition 10.4.*

We see from Proposition 10.4 that, for the case of affine operators T whose linear part is not Hermitian, instead of asking if T is av, we can ask if T is pc; since B will almost certainly be diagonalizable, we can answer this question by examining the eigenvalues of B.

Unlike the class of averaged operators, the class of paracontractive operators is not necessarily closed to finite products, unless those factor operators have a common fixed point.

10.9.3 The Elsner-Koltracht-Neumann Theorem

Our interest in paracontractions is due to the Elsner-Koltracht-Neumann (EKN) Theorem [125]:

Theorem 10.6 *Let T be pc with respect to some vector norm. If T has fixed points, then the sequence $\{T^k x^0\}$ converges to a fixed point of T, for all starting vectors x^0.*

We follow the development in [125].

Theorem 10.7 *Suppose that there is a vector norm on $\mathbb{R}^J$, with respect to which each T_i is a pc operator, for $i = 1, ..., I$, and that $F = \cap_{i=1}^{I} \text{Fix}(T_i)$ is not empty. For $k = 0, 1, ...$, let $i(k) = k(\text{mod}\, I) + 1$, and $x^{k+1} = T_{i(k)} x^k$. The sequence $\{x^k\}$ converges to a member of F, for every starting vector x^0.*

Proof: Let $y \in F$. Then, for $k = 0, 1, ...$,

$$||x^{k+1} - y|| = ||T_{i(k)} x^k - y|| \leq ||x^k - y||, \tag{10.45}$$

so that the sequence $\{||x^k - y||\}$ is decreasing; let $d \geq 0$ be its limit. Since the sequence $\{x^k\}$ is bounded, we select an arbitrary cluster point, x^*. Then $d = ||x^* - y||$, from which we can conclude that

$$||T_i x^* - y|| = ||x^* - y||, \tag{10.46}$$

and $T_i x^* = x^*$, for $i = 1, ..., I$; therefore, $x^* \in F$. Replacing y, an arbitrary member of F, with x^*, we have that $||x^k - x^*||$ is decreasing. But, a subsequence converges to zero, so the whole sequence must converge to zero. This completes the proof. ∎

Corollary 10.4 *If T is pc with respect to some vector norm, and T has fixed points, then the iterative sequence $\{T^k x^0\}$ converges to a fixed point of T, for every starting vector x^0.*

Corollary 10.5 *If $T = T_I T_{I-1} \cdots T_2 T_1$, and $F = \cap_{i=1}^{I} \text{Fix}\,(T_i)$ is not empty, then $F = \text{Fix}\,(T)$.*

Proof: The sequence $x^{k+1} = T_{i(k)} x^k$ converges to a member of $\text{Fix}\,(T)$, for every x^0. Select x^0 in F. ∎

Corollary 10.6 *The product T of two or more pc operators T_i, $i = 1, ..., I$ is again a pc operator, if $F = \cap_{i=1}^{I} \text{Fix}\,(T_i)$ is not empty.*

Proof: Suppose that for $T = T_I T_{I-1} \cdots T_2 T_1$, and $y \in F = \text{Fix}\,(T)$, we have

$$||Tx - y|| = ||x - y||. \tag{10.47}$$

Then, since

$$||T_I (T_{I-1} \cdots T_1) x - y|| \leq ||T_{I-1} \cdots T_1 x - y||$$

$$\leq ... \leq ||T_1 x - y|| \leq ||x - y||, \tag{10.48}$$

it follows that

$$||T_i x - y|| = ||x - y||, \tag{10.49}$$

and $T_i x = x$, for each i. Therefore, $Tx = x$. ∎

10.10 Applications of the KM Theorem

In this section we sketch briefly several instances in which the KM Theorem can be applied to obtain convergence of iterative algorithms. These methods will be considered in more detail later in the text.

10.10.1 The ART

The ART is an iterative method for solving $Ax = b$, a system of I linear equations in J unknowns. For each $i = 1, ..., I$, we denote by P_i the orthogonal projection onto the hyperplane

$$H_i = \{x | (Ax)_i = b_i\},$$

and let

$$T = P_I P_{I-1} \cdots P_2 P_1.$$

The *full-cycle* ART sequence is defined as $z^m = T^m z^0$, for $m = 1, 2, ...$ The operators P_i are averaged, therefore so is T. If the system has solutions, then T has fixed points, and, by the KM Theorem 10.3, the sequence $\{z^m\}$ converges to a solution of $Ax = b$. Later, we shall prove that the limit is the solution closest to z^0 in the 2-norm.

10.10.2 The CQ Algorithm

Let A be a real I by J matrix, and C and Q given closed convex sets in $\mathbb{R}^J$ and $\mathbb{R}^I$, respectively. The *split feasibility problem* (SFP) is to find x in C with Ax in Q. In [62] the CQ algorithm for solving the SFP was presented, for the real case. It has the iterative step

$$x^{k+1} = P_C(x^k - \gamma A^T (I - P_Q) A x^k), \tag{10.50}$$

where I is the identity operator and $\gamma \in (0, 2/\rho(A^T A))$, for $\rho(A^T A)$ the spectral radius of the matrix $A^T A$, which is also its largest eigenvalue. The CQ algorithm can be extended to the complex case, in which the matrix A

has complex entries, and the sets C and Q are in $\mathbb{C}^J$ and $\mathbb{C}^I$, respectively. The iterative step of the extended CQ algorithm is then

$$x^{k+1} = P_C(x^k - \gamma A^\dagger(I - P_Q)Ax^k). \tag{10.51}$$

The CQ algorithm converges to a solution of the SFP, for any starting vector x^0, whenever the SFP has solutions. When the SFP has no solutions, the CQ algorithm converges to a minimizer of the function

$$f(x) = \frac{1}{2}||P_Q Ax - Ax||_2^2$$

over the set C, provided such constrained minimizers exist [62, 63]. The CQ algorithm employs the relaxation parameter γ in the interval $(0, 2/L)$, where L is the largest eigenvalue of the matrix $A^T A$. Choosing the best relaxation parameter in any algorithm is a nontrivial procedure. Generally speaking, we want to select γ near to $1/L$.

For the real case, the gradient of the function $f(x)$ is

$$\nabla f(x) = A^T(I - P_Q)Ax. \tag{10.52}$$

It is shown in [63] that the operator $\nabla f(x)$ is $\rho(A^T A)$-Lipschitz and therefore is ν-ism for $\nu = 1/\rho(A^T A)$. It follows that the operator $T = P_C(I - \gamma\nabla f)$ is averaged, whenever $0 < \gamma < 2/\rho(A^T A)$. Consequently, the CQ algorithm converges to a minimizer of $f(x)$ over x in C, whenever such minimizers exist.

With $A = I$ and $C = Q$, it follows that the gradient of the function

$$h(x) = \frac{1}{2}\|x - P_C x\|_2^2$$

is $\nabla h(x) = x - P_C x$, and therefore the gradient of the function

$$g(x) = \frac{1}{2}\Big(\|x\|_2^2 - \|x - P_C x_2^2\Big)$$

is $\nabla g(x) = P_C x$.

10.10.3 Landweber's Algorithm

If we select $C = \mathbb{R}^J$ and $Q = \{b\}$, then the CQ algorithm has the iterative step

$$x^{k+1} = x^k - \gamma A^T(Ax^k - b), \tag{10.53}$$

which is the iterative step of Landweber's algorithm. Therefore, Landweber's algorithm converges to a minimizer of the function $f(x) = \|Ax - b\|_2$.

10.10.4 Projected Landweber's Algorithm

Let C be an arbitrary closed convex set in $\mathbb{R}^J$. The *projected* Landweber's algorithm has the iterative step

$$x^{k+1} = P_C(x^k - \gamma A^T(Ax^k - b)). \tag{10.54}$$

This also is a special case of the CQ algorithm, so we know that the projected Landweber algorithm converges to a minimizer, over x in C, of the function $f(x) = \|Ax - b\|_2$, whenever such minimizers exist.

10.10.5 Successive Orthogonal Projection

Let $C_1, ..., C_I$ be closed convex sets in $\mathbb{R}^J$, with non-empty intersection C. The *convex feasibility problem* (CFP) is to find a member of C. Typically, the orthogonal projection onto each C_i is easy to calculate, but the orthogonal projection onto C is not. The objective is to find a member of C, not necessarily the orthogonal projection of x^0, using the orthogonal projections onto each C_i in turn.

For each i let P_i be the orthogonal projection operator for C_i and define $T = P_I P_{I-1} \cdots P_2 P_1$. Then T is averaged. If C is non-empty, then the members of C are the fixed points of T. The *successive orthogonal projection* algorithm (SOP) has the iterative step $x^m = T^m x^0$. According to the KM Theorem 10.3, the SOP sequence converges to a member of C.

Chapter 11

Jacobi and Gauss-Seidel Methods

11.1 Chapter Summary

In this chapter we consider two well known iterative algorithms for solving square systems of linear equations, the Jacobi method and the Gauss-Seidel method. Both these algorithms are easy to describe and to motivate. They both require not only that the system be square, that is, have the same number of unknowns as equations, but satisfy additional constraints needed for convergence.

Linear systems $Ax = b$ need not be square but can be associated with two square systems, $A^\dagger Ax = A^\dagger b$, the so-called *normal equations*, and $AA^\dagger z = b$, sometimes called the *Björck-Elfving equations* [108]. Both the Jacobi and the Gauss-Seidel algorithms can be modified to apply to any square system of linear equations, $Sz = h$. The resulting algorithms, the Jacobi overrelaxation (JOR) and successive overrelaxation (SOR) methods, involve the choice of a parameter. The JOR and SOR will converge for more general classes of matrices, provided that the parameter is appropriately chosen.

When we say that an iterative method is convergent, or converges, under certain conditions, we mean that it converges for any consistent system of the appropriate type, and for any starting vector; any iterative method

will converge if we begin at the right answer. We assume throughout this chapter that A is an I by J matrix.

11.2 The Jacobi and Gauss-Seidel Methods: An Example

Suppose we wish to solve the 3 by 3 system

$$S_{11}z_1 + S_{12}z_2 + S_{13}z_3 = h_1$$

$$S_{21}z_1 + S_{22}z_2 + S_{23}z_3 = h_2$$

$$S_{31}z_1 + S_{32}z_2 + S_{33}z_3 = h_3, \tag{11.1}$$

which we can rewrite as

$$z_1 = S_{11}^{-1}[h_1 - S_{12}z_2 - S_{13}z_3]$$

$$z_2 = S_{22}^{-1}[h_2 - S_{21}z_1 - S_{23}z_3]$$

$$z_3 = S_{33}^{-1}[h_3 - S_{31}z_1 - S_{32}z_2], \tag{11.2}$$

assuming that the diagonal terms S_{mm} are not zero. Let $z^0 = (z_1^0, z_2^0, z_3^0)^T$ be an initial guess for the solution. We then insert the entries of z^0 on the right sides and use the left sides to define the entries of the next guess z^1. This is one full cycle of *Jacobi's method*.

The Gauss-Seidel method is similar. Let $z^0 = (z_1^0, z_2^0, z_3^0)^T$ be an initial guess for the solution. We then insert z_2^0 and z_3^0 on the right side of the first equation, obtaining a new value z_1^1 on the left side. We then insert z_3^0 and z_1^1 on the right side of the second equation, obtaining a new value z_2^1 on the left. Finally, we insert z_1^1 and z_2^1 into the right side of the third equation, obtaining a new z_3^1 on the left side. This is one full cycle of the *Gauss-Seidel* (GS) method.

11.3 Splitting Methods

The Jacobi and the Gauss-Seidel methods are particular cases of a more general approach known as *splitting methods*. Splitting methods apply to

square systems of linear equations. Let S be an arbitrary N by N square matrix, written as $S = M - K$. Then the linear system of equations $Sz = h$ is equivalent to $Mz = Kz + h$. If M is invertible, then we can also write $z = M^{-1}Kz + M^{-1}h$. This last equation suggests a class of iterative methods for solving $Sz = h$ known as *splitting methods*. The idea is to select a matrix M so that the equation

$$Mz^{k+1} = Kz^k + h \tag{11.3}$$

can be easily solved to get z^{k+1}; in the Jacobi method M is diagonal, and in the Gauss-Seidel method, M is triangular. Then we write

$$z^{k+1} = M^{-1}Kz^k + M^{-1}h. \tag{11.4}$$

From $K = M - S$, we can write Equation (11.4) as

$$z^{k+1} = z^k + M^{-1}(h - Sz^k). \tag{11.5}$$

Suppose that S is invertible and $\hat{z}$ is the unique solution of $Sz = h$. The error we make at the k-th step is $e^k = \hat{z} - z^k$, so that

$$e^{k+1} = M^{-1}Ke^k.$$

We want the error to decrease with each step, which means that we should seek M and K so that $||M^{-1}K|| < 1$. If S is not invertible and there are multiple solutions of $Sz = h$, then we do not want $M^{-1}K$ to be a strict contraction, but only av or pc. The operator T defined by

$$Tz = M^{-1}Kz + M^{-1}h = Bz + d \tag{11.6}$$

is an affine linear operator and will be a pc or av operator whenever $B = M^{-1}K$ is.

It follows from our previous discussion concerning linear av operators that, if $B = B^\dagger$ is Hermitian, then B is av if and only if

$$-1 < \lambda \leq 1, \tag{11.7}$$

for all (necessarily real) eigenvalues λ of B.

In general, though, the matrix $B = M^{-1}K$ will not be Hermitian, and deciding if such a non-Hermitian matrix is av is not a simple matter. We do know that, if B is av, so is $B^\dagger$; the matrix B is a convex combination of the identity and a non-expansive matrix N, so $B^\dagger$ is a convex combination of the identity and $N^\dagger$, which is also non-expansive, since $\|N^\dagger\| = \|N\| \leq 1$. Consequently, the Hermitian matrix $Q = \frac{1}{2}(B + B^\dagger)$ is also av. Therefore, $I - Q = \frac{1}{2}(M^{-1}S + (M^{-1}S)^\dagger)$ is ism, and so is non-negative definite. We have $-1 < \lambda \leq 1$, for any eigenvalue λ of Q.

Alternatively, we can use the EKN Theorem 10.6. According to that

theorem, if B has a basis of eigenvectors, and $|\lambda| < 1$ for all eigenvalues λ of B that are not equal to one, then $\{z^k\}$ will converge to a solution of $Sz = h$, whenever solutions exist.

In what follows we shall write an arbitrary square matrix S as

$$S = L + D + U, \tag{11.8}$$

where L is the strictly lower triangular part of S, D the diagonal part, and U the strictly upper triangular part. When $S = H$ is Hermitian, we have

$$H = L + D + L^{\dagger}. \tag{11.9}$$

We list now several examples of iterative algorithms obtained by the splitting method. In the remainder of the chapter we discuss these methods in more detail.

11.4 Some Examples of Splitting Methods

As we shall now see, the Jacobi and Gauss-Seidel methods, as well as their overrelaxed versions, JOR and SOR, are splitting methods.

Jacobi's Method: Jacobi's method uses $M = D$ and $K = -L - U$, under the assumption that D is invertible. The matrix B is

$$B = M^{-1}K = -D^{-1}(L + U). \tag{11.10}$$

The Gauss-Seidel Method: The Gauss-Seidel (GS) method uses the splitting $M = D + L$, so that the matrix B is

$$B = I - (D + L)^{-1}S. \tag{11.11}$$

The Jacobi Overrelaxation Method (JOR): The JOR uses the splitting

$$M = \frac{1}{\omega}D \tag{11.12}$$

and

$$K = M - S = (\frac{1}{\omega} - 1)D - L - U. \tag{11.13}$$

The matrix B is

$$B = M^{-1}K = (I - \omega D^{-1}S). \tag{11.14}$$

The Successive Overrelaxation Method (SOR): The SOR uses the splitting $M = (\frac{1}{\omega}D + L)$, so that

$$B = M^{-1}K = (D + \omega L)^{-1}[(1-\omega)D - \omega U] \tag{11.15}$$

or

$$B = I - \omega(D + \omega L)^{-1}S, \tag{11.16}$$

or

$$B = (I + \omega D^{-1}L)^{-1}[(1-\omega)I - \omega D^{-1}U]. \tag{11.17}$$

11.5 Jacobi's Algorithm and JOR

The matrix B in Equation (11.10) is not generally av and the Jacobi iterative scheme will not converge, in general. Additional conditions need to be imposed on S in order to guarantee convergence. One such condition is that S be strictly diagonally dominant. In that case, all the eigenvalues of $B = M^{-1}K$ can be shown to lie inside the unit circle of the complex plane, so that $\rho(B) < 1$. It follows from Lemma 6.3 that B is sc with respect to some vector norm, and the Jacobi iteration converges. If, in addition, S is Hermitian, the eigenvalues of B are in the interval $(-1, 1)$, and so B is sc with respect to the Euclidean norm.

Alternatively, one has the *Jacobi overrelaxation* (JOR) method, which is essentially a special case of the Landweber algorithm and involves an arbitrary parameter.

For S an N by N matrix, Jacobi's method can be written as

$$z_m^{\text{new}} = S_{mm}^{-1}[h_m - \sum_{j \neq m} S_{mj} z_j^{\text{old}}], \tag{11.18}$$

for $m = 1, ..., N$. With D the invertible diagonal matrix with entries $D_{mm} = S_{mm}$ we can write one cycle of Jacobi's method as

$$z^{\text{new}} = z^{\text{old}} + D^{-1}(h - Sz^{\text{old}}). \tag{11.19}$$

The *Jacobi overrelaxation* (JOR) method has the following full-cycle iterative step:

$$z^{\text{new}} = z^{\text{old}} + \omega D^{-1}(h - Sz^{\text{old}}); \tag{11.20}$$

choosing $\omega = 1$ we get the Jacobi method. Convergence of the JOR iteration will depend, of course, on properties of S and on the choice of ω. When $S = Q$, where Q is Hermitian and nonnegative-definite, for example, $S = A^\dagger A$ or $S = AA^\dagger$, we can say more. Note that such Q can always be written in the form $Q = AA^\dagger$ or $Q = A^\dagger A$, for appropriately chosen A.

11.5.1 The JOR in the Nonnegative-definite Case

When $S = Q$ is nonnegative-definite and the system $Qz = h$ is consistent the JOR converges to a solution for any $\omega \in (0, 2/\rho(D^{-1/2}QD^{-1/2}))$, where $\rho(Q)$ denotes the largest eigenvalue of the nonnegative-definite matrix Q. For nonnegative-definite Q, the convergence of the JOR method is implied by the KM Theorem 10.3, since the JOR is equivalent to Landweber's algorithm in these cases. To see this, we rewrite Equation (11.20) as

$$v^{\text{new}} = v^{\text{old}} + \omega G^\dagger (f - Gv^{\text{old}}),$$

where $v = D^{1/2}z$,

$$G^\dagger G = D^{-1/2}QD^{-1/2},$$

and

$$G^\dagger f = D^{-1/2}h.$$

The JOR method, as applied to $Qz = AA^\dagger z = b$, is equivalent to the Landweber iterative method for $Ax = b$.

Ex. 11.1 *Show that the system $AA^\dagger z = b$ has solutions whenever the system $Ax = b$ has solutions.*

Lemma 11.1 *If $\{z^k\}$ is the sequence obtained from the JOR, then the sequence $\{A^\dagger z^k\}$ is the sequence obtained by applying the Landweber algorithm to the system $D^{-1/2}Ax = D^{-1/2}b$, where D is the diagonal part of the matrix $Q = AA^\dagger$.*

If we select $\omega = 1/I$ we obtain the Cimmino method. Since the trace of the matrix $D^{-1/2}QD^{-1/2}$ equals I, which then is the sum of its eigenvalues, all of which are non-negative, we know that $\omega = 1/I$ is less than two over the largest eigenvalue of the matrix $D^{-1/2}QD^{-1/2}$ and so this choice of ω is acceptable and the Cimmino algorithm converges whenever there are solutions of $Ax = b$. In fact, it can be shown that Cimmino's method converges to a least squares approximate solution generally.

Similarly, the JOR method applied to the system $A^\dagger Ax = A^\dagger b$ is equivalent to the Landweber algorithm, applied to the system $Ax = b$.

Ex. 11.2 *Show that, if $\{z^k\}$ is the sequence obtained from the JOR, then the sequence $\{D^{1/2}z^k\}$ is the sequence obtained by applying the Landweber algorithm to the system $AD^{-1/2}x = b$, where D is the diagonal part of the matrix $S = A^\dagger A$.*

11.6 The Gauss-Seidel Algorithm and SOR

In general, the full-cycle iterative step of the Gauss-Seidel method is the following:

$$z^{\text{new}} = z^{\text{old}} + (D + L)^{-1}(h - Sz^{\text{old}}), \tag{11.21}$$

where $S = D + L + U$ is the decomposition of the square matrix S into its diagonal, lower triangular and upper triangular diagonal parts. The GS method does not converge without restrictions on the matrix S. As with the Jacobi method, strict diagonal dominance is a sufficient condition.

11.6.1 The Nonnegative-Definite Case

Now we consider the square system $Qz = h$, assuming that $Q = L + D + L^\dagger$ is Hermitian and nonnegative-definite, so that $x^\dagger Qx \geq 0$, for all x. It is easily shown that all the entries of D are nonnegative. We assume that all the diagonal entries of D are positive, so that $D + L$ is invertible. The Gauss-Seidel iterative step is $z^{k+1} = Tz^k$, where T is the affine linear operator given by $Tz = Bz+d$, for $B = -(D+L)^{-1}L^\dagger$ and $d = (D+L)^{-1}h$.

Proposition 11.1 *Let λ be an eigenvalue of B that is not equal to one. Then $|\lambda| < 1$.*

If B is diagonalizable, then there is a norm with respect to which T is paracontractive, so, by the EKN Theorem 10.6, the GS iteration converges to a solution of $Qz = h$, whenever solutions exist.

Proof of Proposition (11.1): Let $Bv = \lambda v$, for v nonzero. Then $-Bv = (D + L)^{-1}L^\dagger v = -\lambda v$, so that

$$L^\dagger v = -\lambda(D + L)v, \tag{11.22}$$

and

$$Lv = -\overline{\lambda}(D + L)^\dagger v. \tag{11.23}$$

Therefore,

$$v^\dagger L^\dagger v = -\lambda v^\dagger (D + L)v. \tag{11.24}$$

Adding $v^\dagger(D + L)v$ to both sides, we get

$$v^\dagger Qv = (1 - \lambda)v^\dagger(D + L)v. \tag{11.25}$$

Since the left side of the equation is real, so is the right side. Therefore

$$\begin{aligned}(1-\overline{\lambda})(D+L)^{\dagger}v &= (1-\lambda)v^{\dagger}(D+L)v \\ &= (1-\lambda)v^{\dagger}Dv + (1-\lambda)v^{\dagger}Lv \\ &= (1-\lambda)v^{\dagger}Dv - (1-\lambda)\overline{\lambda}v^{\dagger}(D+L)^{\dagger}v. \end{aligned} \tag{11.26}$$

So we have

$$[(1-\overline{\lambda}) + (1-\lambda)\overline{\lambda}]v^{\dagger}(D+L)^{\dagger}v = (1-\lambda)v^{\dagger}Dv, \tag{11.27}$$

or

$$(1-|\lambda|^2)v^{\dagger}(D+L)^{\dagger}v = (1-\lambda)v^{\dagger}Dv. \tag{11.28}$$

Multiplying by $(1-\overline{\lambda})$ on both sides, we get, on the left side,

$$(1-|\lambda|^2)v^{\dagger}(D+L)^{\dagger}v - (1-|\lambda|^2)\overline{\lambda}v^{\dagger}(D+L)^{\dagger}v, \tag{11.29}$$

which is equal to

$$(1-|\lambda|^2)v^{\dagger}(D+L)^{\dagger}v + (1-|\lambda|^2)v^{\dagger}Lv, \tag{11.30}$$

and, on the right side, we get

$$|1-\lambda|^2 v^{\dagger}Dv. \tag{11.31}$$

Consequently, we have

$$(1-|\lambda|^2)v^{\dagger}Qv = |1-\lambda|^2 v^{\dagger}Dv. \tag{11.32}$$

Since $v^{\dagger}Qv \geq 0$ and $v^{\dagger}Dv > 0$, it follows that $1-|\lambda|^2 \geq 0$. If $|\lambda| = 1$, then $|1-\lambda|^2 = 0$, so that $\lambda = 1$. This completes the proof. ∎

Note that $\lambda = 1$ if and only if $Qv = 0$. Therefore, if Q is invertible, the affine linear operator T is a strict contraction, and the GS iteration converges to the unique solution of $Qz = h$.

11.6.2 The GS Algorithm as ART

We show now that the GS algorithm, when applied to the system $Qz = AA^{\dagger}z = b$, is equivalent to the ART algorithm, applied to $Ax = b$. Let $AA^{\dagger} = Q = L + D + L^{\dagger}$.

It is convenient now to consider separately each sub-iteration step of the GS algorithm. For $m = 0, 1, ...$ and $i = m(\mathrm{mod}\, I) + 1$, we denote by z^{m+1} the vector whose entries are

$$z_i^{m+1} = D_{ii}^{-1}\Big(b_i - (Qz^m)_i + Q_{ii}z_i^m\Big),$$

and $z_n^{m+1} = z_n^m$, for $n \neq i$. Therefore, we can write

$$z_i^{m+1} - z_i^m = D_{ii}^{-1}(b_i - (AA^\dagger z^m)_i).$$

Now let $x^m = A^\dagger z^m$ for each m. Then we have

$$x_j^{m+1} = (A^\dagger z^{m+1})_j = (A^\dagger z^m)_j + \overline{A_{ij}} D_{ii}^{-1}(b_i - (Ax^m)_i),$$

which is one step of the ART algorithm, applied to the system $Ax = b$. Note that

$$D_{ii} = \sum_{j=1}^{J} |A_{ij}|^2.$$

From this, we can conclude that if $\{z^k\}$ is the sequence produced by one step of the GS algorithm, applied to the system $AA^\dagger z = b$, then $\{x^k = A^\dagger z^k\}$ is the sequence produced by one full cycle of the ART algorithm, applied to the system $Ax = b$. Since we know that the ART algorithm converges whenever $Ax = b$ is consistent, we know now that the GS algorithm, applied to the system $AA^\dagger z = b$, converges whenever $Ax = b$ is consistent. So once again we have shown that when $S = Q$ is Hermitian and non-negative definite, the GS method converges whenever there are solutions of $Qz = h$.

11.6.3 Successive Overrelaxation

The *successive overrelaxation* (SOR) method has the following full-cycle iterative step:

$$z^{\text{new}} = z^{\text{old}} + (\omega^{-1}D + L)^{-1}(h - Sz^{\text{old}}); \tag{11.33}$$

the choice of $\omega = 1$ gives the GS method. Convergence of the SOR iteration will depend, of course, on properties of S and on the choice of ω.

Using the form

$$B = (D + \omega L)^{-1}[(1 - \omega)D - \omega U] \tag{11.34}$$

we can show that

$$|\det(B)| = |1 - \omega|^N. \tag{11.35}$$

From this and the fact that the determinant of B is the product of its eigenvalues, we conclude that $\rho(B) > 1$ if $\omega < 0$ or $\omega > 2$. When $S = Q$ is Hermitian and nonnegative-definite, we can say more.

11.6.4 The SOR for Nonnegative-Definite Q

When Q is nonnegative-definite and the system $Qz = h$ is consistent the SOR converges to a solution for any $\omega \in (0, 2)$. This follows from the convergence of the ART algorithm, since, for such Q, the SOR is equivalent to the ART, as we now show.

Now we write $Q = AA^\dagger$ and consider the SOR method applied to the Björck-Elfving equations $AA^\dagger z = b$. Rather than count a full cycle as one iteration, we now count as a single step the calculation of a single new entry. Therefore, for $k = 0, 1, ...$ the $k+1$-st step replaces the value z_i^k only, where $i = k(\mathrm{mod}\, I) + 1$. We have

$$z_i^{k+1} = (1-\omega)z_i^k + \omega D_{ii}^{-1}(b_i - \sum_{n=1}^{i-1} Q_{in}z_n^k - \sum_{n=i+1}^{I} Q_{in}z_n^k) \tag{11.36}$$

and $z_n^{k+1} = z_n^k$ for $n \neq i$. Now we calculate $x^{k+1} = A^\dagger z^{k+1}$:

$$x_j^{k+1} = x_j^k + \omega D_{ii}^{-1}\overline{A_{ij}}(b_i - (Ax^k)_i). \tag{11.37}$$

This is one step of the relaxed *algebraic reconstruction technique* (ART) applied to the original system of equations $Ax = b$. The relaxed ART converges to a solution, when solutions exist, for any $\omega \in (0, 2)$.

When $Ax = b$ is consistent, so is $AA^\dagger z = b$. We consider now the case in which $Q = AA^\dagger$ is invertible. Since the relaxed ART sequence $\{x^k = A^\dagger z^k\}$ converges to a solution x^∞, for any $\omega \in (0, 2)$, the sequence $\{AA^\dagger z^k\}$ converges to b. Since $Q = AA^\dagger$ is invertible, the SOR sequence $\{z^k\}$ then converges to $Q^{-1}b$.

11.7 Summary

We summarize the basic points of this chapter:

- **1.** Splitting methods for solving $Sz = h$, for square matrix $S = M - K$, involve affine linear operators $Tx = Bx + d$, where $B = M^{-1}K$ and $d = M^{-1}h$;
- **2.** T is av if and only if B is av; if B is Hermitian, then B is av if and only if $-1 < \lambda \leq 1$ for all eigenvalues λ of B;
- **3.** if B is not Hermitian, but is diagonalizable, and $|\lambda| < 1$ unless $\lambda = 1$, then there is a norm for which T is pc;
- **4.** If S is strictly diagonally dominant, then the Jacobi and Gauss-Seidel iterations converge;

- **5.** When $S = Q$ is Hermitian and non-negative definite, Q can be written as either $AA^\dagger$ or as $A^\dagger A$, for appropriately chosen A, and the JOR method is equivalent to Landweber's algorithm for either $D^{-1/2}Ax = D^{-1/2}b$ or $AD^{-1/2}x = b$;

- **6.** When $S = Q$ is Hermitian and non-negative definite, and we write $Q = AA^\dagger$, the SOR method is equivalent to the relaxed ART algorithm for $Ax = b$, and so converges whenever there are solutions, for $0 < \omega < 2$.

Chapter 12

A Tale of Two Algorithms

12.1 Chapter Summary

Although the EMML and SMART algorithms have quite different histories and are not typically considered together, they are closely related, as we shall see [51, 52]. In this chapter we examine these two algorithms in tandem, following [53]. Forging a link between the EMML and SMART led to a better understanding of both of these algorithms and to new results. The proof of convergence of the SMART in the inconsistent case [51] was based on the analogous proof for the EMML [258], while discovery of the faster version of the EMML, the *rescaled block-iterative* EMML (RBI-EMML) [54] came from studying the analogous block-iterative version of SMART [89]. The proofs we give here are elementary and rely mainly on easily established properties of the cross-entropy or Kullback-Leibler distance.

12.2 Notation

Let A be an I by J matrix with entries $A_{ij} \geq 0$, such that, for each $j = 1, ..., J$, we have $s_j = \sum_{i=1}^{I} A_{ij} > 0$. Let $b = (b_1, ..., b_I)^T$ with $b_i > 0$ for each i. We shall assume throughout this chapter that $s_j = 1$ for each j. If this is not the case initially, we replace x_j with $x_j s_j$ and A_{ij} with A_{ij}/s_j; the quantities $(Ax)_i$ are unchanged.

12.3 The Two Algorithms

The algorithms we shall consider are the expectation maximization maximum likelihood method (EMML) and the simultaneous multiplicative algebraic reconstruction technique (SMART). When $b = Ax$ has nonnegative solutions, both algorithms produce such a solution. In general, the EMML gives a nonnegative minimizer of $KL(b, Ax)$, while the SMART minimizes $KL(Ax, b)$ over nonnegative x.

For both algorithms we begin with an arbitrary positive vector x^0. The iterative step for the EMML method is

$$x_j^{k+1} = (x^k)'_j = x_j^k \sum_{i=1}^{I} A_{ij} \frac{b_i}{(Ax^k)_i}. \tag{12.1}$$

The iterative step for the SMART is

$$x_j^{m+1} = (x^m)''_j = x_j^m \exp\Big(\sum_{i=1}^{I} A_{ij} \log \frac{b_i}{(Ax^m)_i}\Big). \tag{12.2}$$

Note that, to avoid confusion, we use k for the iteration number of the EMML and m for the SMART.

12.4 Background

The expectation maximization maximum likelihood method (EMML) has been the subject of much attention in the medical-imaging literature over the past decade. Statisticians like it because it is based on the

well-studied principle of likelihood maximization for parameter estimation. Physicists like it because, unlike its competition, filtered back-projection, it permits the inclusion of sophisticated models of the physical situation. Mathematicians like it because it can be derived from iterative optimization theory. Physicians like it because the images are often better than those produced by other means. No method is perfect, however, and the EMML suffers from sensitivity to noise and slow rate of convergence. Research is ongoing to find faster and less sensitive versions of this algorithm.

Another class of iterative algorithms was introduced into medical imaging by Gordon et al. in [151]. These include the *algebraic reconstruction technique* (ART) and its multiplicative version, MART. These methods were derived by viewing image reconstruction as solving systems of linear equations, possibly subject to constraints, such as positivity. The *simultaneous* MART (SMART) [107, 233] is a variant of MART that uses all the data at each step of the iteration.

12.5 The Kullback-Leibler Distance

The Kullback-Leibler distance $KL(x, z)$ is defined for nonnegative vectors x and z by Equations (24.18) and (24.19). Clearly, the KL distance has the property $KL(cx, cz) = cKL(x, z)$ for all positive scalars c.

Ex. 12.1 *Let* $z_+ = \sum_{j=1}^{J} z_j > 0$. *Then*

$$KL(x, z) = KL(x_+, z_+) + KL(x, (x_+/z_+)z). \tag{12.3}$$

As we shall see, the KL distance mimics the ordinary Euclidean distance in several ways that make it particularly useful in designing optimization algorithms. The following exercise shows that the KL distance does exhibit some behavior not normally associated with a distance.

Ex. 12.2 *Let* x *be in the interval* $(0, 1)$. *Show that*

$$KL(x, 1) + KL(1, x^{-1}) < KL(x, x^{-1}).$$

12.6 The Alternating Minimization Paradigm

For each nonnegative vector x for which $(Ax)_i = \sum_{j=1}^{J} A_{ij}x_j > 0$, let $r(x) = \{r(x)_{ij}\}$ and $q(x) = \{q(x)_{ij}\}$ be the I by J arrays with entries

$$r(x)_{ij} = x_j A_{ij} \frac{b_i}{(Ax)_i}$$

and

$$q(x)_{ij} = x_j A_{ij}.$$

The KL distances

$$KL(r(x), q(z)) = \sum_{i=1}^{I} \sum_{j=1}^{J} KL(r(x)_{ij}, q(z)_{ij})$$

and

$$KL(q(x), r(z)) = \sum_{i=1}^{I} \sum_{j=1}^{J} KL(q(x)_{ij}, r(z)_{ij})$$

will play important roles in the discussion that follows. Note that if there is nonnegative x with $r(x) = q(x)$ then $b = Ax$.

12.6.1 Some Pythagorean Identities Involving the KL Distance

The iterative algorithms we discuss in this chapter are derived using the principle of *alternating minimization*, according to which the distances $KL(r(x), q(z))$ and $KL(q(x), r(z))$ are minimized, first with respect to the variable x and then with respect to the variable z. Although the KL distance is not Euclidean, and, in particular, not even symmetric, there are analogues of Pythagoras' theorem that play important roles in the convergence proofs.

Ex. 12.3 *Establish the following Pythagorean identities:*

$$KL(r(x), q(z)) = KL(r(z), q(z)) + KL(r(x), r(z)); \tag{12.4}$$

$$KL(r(x), q(z)) = KL(r(x), q(x')) + KL(x', z), \tag{12.5}$$

for

$$x'_j = x_j \sum_{i=1}^{I} A_{ij} \frac{b_i}{(Ax)_i}; \tag{12.6}$$

$$KL(q(x), r(z)) = KL(q(x), r(x)) + KL(x, z) - KL(Ax, Az); \quad (12.7)$$

$$KL(q(x), r(z)) = KL(q(z''), r(z)) + KL(x, z''), \quad (12.8)$$

for

$$z''_j = z_j \exp(\sum_{i=1}^{I} A_{ij} \log \frac{b_i}{(Az)_i}). \quad (12.9)$$

Note that it follows from Equation (12.3) that $KL(x, z) - KL(Ax, Az) \geq 0$.

12.6.2 Convergence of the SMART and EMML

We shall prove convergence of the SMART and EMML algorithms through a series of exercises.

Ex. 12.4 *Show that, for* $\{x^k\}$ *given by Equation (12.1),* $\{KL(b, Ax^k)\}$ *is decreasing and* $\{KL(x^{k+1}, x^k)\} \to 0$. *Show that, for* $\{x^m\}$ *given by Equation (12.2),* $\{KL(Ax^m, b)\}$ *is decreasing and* $\{KL(x^m, x^{m+1})\} \to 0$. *Hint: Use* $KL(r(x), q(x)) = KL(b, Ax)$, $KL(q(x), r(x)) = KL(Ax, b)$, *and the Pythagorean identities.*

Ex. 12.5 *Show that the EMML sequence* $\{x^k\}$ *is bounded by showing*

$$\sum_{j=1}^{J} x_j^{k+1} = \sum_{i=1}^{I} b_i.$$

Show that the SMART sequence $\{x^m\}$ *is bounded by showing that*

$$\sum_{j=1}^{J} x_j^{m+1} \leq \sum_{i=1}^{I} b_i.$$

Ex. 12.6 *Show that* $(x^*)' = x^*$ *for any cluster point* x^* *of the EMML sequence* $\{x^k\}$ *and that* $(x^*)'' = x^*$ *for any cluster point* x^* *of the SMART sequence* $\{x^m\}$. *Hint: Use* $\{KL(x^{k+1}, x^k)\} \to 0$ *and* $\{KL(x^m, x^{m+1})\} \to 0$.

Ex. 12.7 *Let* $\hat{x}$ *and* $\tilde{x}$ *minimize* $KL(b, Ax)$ *and* $KL(Ax, b)$, *respectively, over all* $x \geq 0$. *Then,* $(\hat{x})' = \hat{x}$ *and* $(\tilde{x})'' = \tilde{x}$. *Hint: Apply Pythagorean identities to* $KL(r(\hat{x}), q(\hat{x}))$ *and* $KL(q(\tilde{x}), r(\tilde{x}))$.

Note that, because of convexity properties of the KL distance, even if the minimizers $\hat{x}$ and $\tilde{x}$ are not unique, the vectors $A\hat{x}$ and $A\tilde{x}$ are unique.

Ex. 12.8 *For the EMML sequence $\{x^k\}$ with cluster point x^* and $\hat{x}$ as defined previously, we have the double inequality*

$$KL(\hat{x}, x^k) \geq KL(r(\hat{x}), r(x^k)) \geq KL(\hat{x}, x^{k+1}), \tag{12.10}$$

from which we conclude that the sequence $\{KL(\hat{x}, x^k)\}$ is decreasing and $KL(\hat{x}, x^*) < +\infty$. Hint: For the first inequality calculate $KL(r(\hat{x}), q(x^k))$ in two ways. For the second one, use $(x)'_j = \sum_{i=1}^{I} r(x)_{ij}$ and Exercise 12.1.

Ex. 12.9 *Show that, for the SMART sequence $\{x^m\}$ with cluster point x^* and $\tilde{x}$ as defined previously, we have*

$$KL(\tilde{x}, x^m) - KL(\tilde{x}, x^{m+1}) = KL(Ax^{m+1}, b) - KL(A\tilde{x}, b) +$$

$$KL(A\tilde{x}, Ax^m) + KL(x^{m+1}, x^m) - KL(Ax^{m+1}, Ax^m), \tag{12.11}$$

and so $KL(A\tilde{x}, Ax^*) = 0$, the sequence $\{KL(\tilde{x}, x^m)\}$ is decreasing and $KL(\tilde{x}, x^*) < +\infty$. Hint: Expand $KL(q(\tilde{x}), r(x^m))$ using the Pythagorean identities.

Ex. 12.10 *For x^* a cluster point of the EMML sequence $\{x^k\}$ we have $KL(b, Ax^*) = KL(b, A\hat{x})$. Therefore, x^* is a nonnegative minimizer of $KL(b, Ax)$. Consequently, the sequence $\{KL(x^*, x^k)\}$ converges to zero, and so $\{x^k\} \to x^*$. Hint: Use the double inequality of Equation (12.10) and $KL(r(\hat{x}), q(x^*))$.*

Ex. 12.11 *For x^* a cluster point of the SMART sequence $\{x^m\}$ we have $KL(Ax^*, b) = KL(A\tilde{x}, b)$. Therefore, x^* is a nonnegative minimizer of $KL(Ax, b)$. Consequently, the sequence $\{KL(x^*, x^m)\}$ converges to zero, and so $\{x^m\} \to x^*$. Moreover,*

$$KL(\tilde{x}, x^0) \geq KL(x^*, x^0)$$

for all $\tilde{x}$ as before. Hints: Use Exercise 12.9. For the final assertion use the fact that the difference $KL(\tilde{x}, x^m) - KL(\tilde{x}, x^{m+1})$ is independent of the choice of $\tilde{x}$, since it depends only on $Ax^* = A\tilde{x}$. Now sum over the index m.

Both the EMML and the SMART algorithms are slow to converge. For that reason attention has shifted, in recent years, to *block-iterative* versions of these algorithms. We take up that topic in a later chapter.

12.7 Sequential Optimization

As we have seen, both the SMART and the EMML algorithms can be developed using the paradigm of alternating minimization. They can also be obtained through *sequential minimization*. For details concerning sequential optimization see [136, 70, 67, 68].

12.7.1 Sequential Unconstrained Optimization

A common, but often difficult, problem in optimization is the minimization or maximization of a function $f(x)$ subject to constraints on the vector variable x. In the sequential unconstrained optimization approach, at the kth step we add to the function $f(x)$ an auxiliary function $g_k(x)$ and optimize

$$G_k(x) = f(x) + g_k(x),$$

to get $x = x^k$ as the result. The $g_k(x)$ are chosen to enforce the constraints, so that each x^k is feasible.

12.7.2 An Example

Suppose that we wish to minimize the function $f(u, v) = u^2+v^2$, subject to the constraint that $u + v \geq 1$. For each $k = 1, 2, ...$ we minimize the function

$$G_k(u, v) = f(u, v) - \frac{1}{k}\log(u + v - 1).$$

In the terminology of sequential optimization, we have added a logarithmic barrier function to $f(x)$. Setting the gradient of $G_k(u, v)$ equal to zero, we find that

$$u_k = v_k = \frac{1}{4} + \frac{1}{4}\sqrt{1 + \frac{4}{k}}.$$

Clearly, $u_k + v_k \geq 1$ and, as $k \to +\infty$, (u_k, v_k) approaches the limit $(\frac{1}{2}, \frac{1}{2})$, which is the answer to the constrained minimization problem.

This method can also be used to facilitate computation. In such cases, we select $g_k(x)$ so that the $x = x^k$ optimizing $G_k(x)$ can be found in closed form. The SMART and EMML algorithms are special cases of this approach to sequential optimization.

12.7.3 The SMART Algorithm

In the case of the SMART algorithm, the function to be minimized is $f(x) = KL(Ax, b)$. The auxiliary function is

$$g_k(x) = KL(x, x^{k-1}) - KL(Ax, Ax^{k-1}) \geq 0.$$

Using the Pythagorean identities, we can show that

$$G_k(x) = KL(q(x), r(x^{k-1})) = f(x) + g_k(x).$$

12.7.4 The EMML Algorithm

In the case of the EMML algorithm, the function to be minimized is $f(x) = KL(b, Ax)$. The auxiliary function is

$$g_k(x) = KL(r(x^{k-1}), r(x)) \geq 0.$$

Using the Pythagorean identities, we can show that

$$G_k(x) = KL(r(x^{k-1}), q(x)) = f(x) + g_k(x).$$

Chapter 13

Block-Iterative Methods I

13.1 Chapter Summary

Both the EMML and the SMART algorithms can be slow to converge. These methods are *simultaneous methods*, in which all the equations are employed at each step of the iteration. Block-iterative methods, in which only some of the equations are used at each step, often converge faster than their simultaneous cousins. In addition, the blocks can be designed to take advantage of the manner in which the computer stores and retrieves data.

13.2 Recalling the MART Algorithm

The MART algorithm uses only one equation at a time. For $k = 0, 1, ...$, we let $i = k(\text{mod}\, I) + 1$ and

$$x_j^{k+1} = x_j^k \Big(\frac{b_i}{(Ax^k)_i} \Big)^{A_{ij} m_i^{-1}}. \tag{13.1}$$

The MART converges to the non-negative solution of $Ax = b$ for which $KL(x, x^0)$ is minimized, whenever such solutions exist, provided that we select m_i so that $A_{ij} \leq m_i$, for all j. Here we shall choose $m_i = \max\{A_{ij} | j = 1, 2, ..., J\}$.

13.3 The EMML and the SMART Algorithms

We recall the formulas for the iterative step of the EMML and the SMART.

13.3.1 The EMML Algorithm

The iterative step for the EMML algorithm is

$$x_j^{k+1} = x_j^k s_j^{-1} \sum_{i=1}^{I} A_{ij} \frac{b_i}{(Ax^k)_i}, \tag{13.2}$$

where $s_j = \sum_{i=1}^{I} A_{ij}$. The iterative step can also be written as

$$x_j^{k+1} = \sum_{i=1}^{I} (s_j^{-1} A_{ij}) \Big(x_j^k \frac{b_i}{(Ax^k)_i} \Big), \tag{13.3}$$

which shows that x_j^{k+1} is the weighted arithmetic mean of the terms $x_j^k \frac{b_i}{(Ax^k)_i}$.

13.3.2 The SMART Algorithm

The iterative step for the SMART algorithm is

$$x_j^{k+1} = x_j^k \exp\Big(s_j^{-1} \sum_{i=1}^{I} A_{ij} \log\Big(\frac{b_i}{(Ax^k)_i} \Big) \Big). \tag{13.4}$$

The iterative step can also be written as

$$x_j^{k+1} = \prod_{i=1}^{I} \Big(x_j^k \frac{b_i}{(Ax^k)_i} \Big)^{s_j^{-1} A_{ij}}, \tag{13.5}$$

which shows that x_j^{k+1} is the weighted geometric mean of the terms $x_j^k \frac{b_i}{(Ax^k)_i}$. In a later section we shall look more closely at these terms.

13.4 Block-Iterative Methods

The term *block-iterative methods* refers to algorithms in which only some of the equations, those in the current block, are used at each step of the iteration. We denote by B_n, $n = 1, ..., N$, the nth block; each B_n is a subset of the index set $\{i = 1, ..., I\}$. The MART is an example of such a block-iterative method; there are $N = I$ blocks, each block containing only one value of the index i. For simplicity, we say that $B_i = \{i\}$, for each i. Once we know x^k, we compute $i = k(\mathrm{mod}\, I) + 1$ and use only the ith equation to compute x^{k+1}.

13.4.1 Block-Iterative SMART

More general block-iterative versions of the SMART algorithm have been known since the work of Darroch and Ratcliff [107], and were treated in detail in [89]. The iterative step of the block-iterative SMART (BI-SMART) algorithm is

$$x_j^{k+1} = x_j^k \exp\Big(m_n^{-1} \sum_{i \in B_n}^{I} A_{ij} \log\Big(\frac{b_i}{(Ax^k)_i}\Big)\Big). \tag{13.6}$$

The BI-SMART converges to the non-negative solution of $Ax = b$ for which $KL(x, x^0)$ is minimized, whenever such solutions exist, provided that $s_{nj} \leq m_n$, where $s_{nj} = \sum_{i \in B_n} A_{ij}$ and $n = k(\mathrm{mod}\, N) + 1$. Here we shall choose $m_n = \max\{s_{nj} | j = 1, 2, ..., J\}$; the BI-SMART with this choice of the parameter m_n is called the *rescaled block-iterative* SMART (RBI-SMART) [54].

13.4.2 Seeking a Block-Iterative EMML

In contrast to the SMART, block-iterative versions of the EMML did not appear in the early literature on this algorithm. The first paper that I am aware of that suggested the use of blocks for the EMML, but without explicit formulas, is the 1990 paper by Holte, Schmidlin *et al.* [170]. Somewhat later, Hudson, Hutton and Larkin [171, 172] discovered what they called the *ordered-subset* (OSEM) variation of the EMML algorithm.

The iterative step of the OSEM algorithm is

$$x_j^{k+1} = x_j^k s_{nj}^{-1} \sum_{i \in B_n} A_{ij} \Big(\frac{b_i}{(Ax^k)_i}\Big). \tag{13.7}$$

It is identical with that of the EMML in Equation (13.2), except that each sum is taken only over the i in the current block B_n.

Although the OSEM often produces usable medical images from tomographic data in much less time than required by the EMML algorithm, there are theoretical problems with OSEM that suggested that OSEM may not be the correct block-iterative version of EMML. First, in order to prove that OSEM converges to a non-negative solution of $Ax = b$, when such solutions exist, we need to assume that the *generalized subset-balance* condition holds: we need

$$s_{nj} = \sum_{i \in B_n} A_{ij} = t_n r_j,$$

for some constants t_n and r_j. Second, if we use the OSEM formula for the case of $N = I$, as in MART, we find that

$$x_j^{k+1} = x_j^k \Big(\frac{b_i}{(Ax^k)_i} \Big),$$

so that each x^{k+1} is simply a scalar multiple of the starting vector x^0; obviously, this is not the proper analog of the MART.

13.4.3 The BI-EMML Algorithm

The problem then is how to define block-iterative versions of the EMML that converge to a non-negative solution whenever there are such solutions, and which give a useful analog of the MART algorithm. To see how to do this, it is helpful to return to the EMML, SMART and MART.

We saw previously that in the SMART, the next iterate x_j^{k+1} is the weighted geometric mean of the terms $x_j^k \Big(\frac{b_i}{(Ax^k)_i} \Big)$, while that of the EMML is the weighted arithmetic mean of the same terms. The MART is also a weighted geometric mean of the single term $x_j^k \Big(\frac{b_i}{(Ax^k)_i} \Big)$ and x_j^k itself; we can write Equation (13.1) as

$$x_j^{k+1} = \Big(x_j^k \Big)^{1 - A_{ij} m_i^{-1}} \Big(x_j^k \frac{b_i}{(Ax^k)_i} \Big)^{A_{ij} m_i^{-1}}. \tag{13.8}$$

This suggests that when we do not use all the equations, we must use x_j^k itself as one of the terms in the weighted geometric or arithmetic mean, which is a form of *relaxation*.

We become more convinced that relaxation is the right idea when we notice that the BI-SMART can be written as

$$x_j^{k+1} = (x_j^k)^{1 - m_n^{-1} s_{nj}} \prod_{i \in B_n} \Big(x_j^k \frac{b_i}{(Ax^k)_i} \Big)^{A_{ij} m_n^{-1}}; \tag{13.9}$$

this tells us that x_j^{k+1} is a weighted geometric mean of x_j^k itself and the terms $x_j^k \Big(\frac{b_i}{(Ax^k)_i} \Big)$ for $i \in B_n$.

Now it becomes clearer how to define the block-iterative EMML algorithms; we must use the weighted arithmetic mean of x_j^k itself and the terms $x_j^k\Big(\frac{b_i}{(Ax^k)_i}\Big)$ for $i \in B_n$. The resulting BI-EMML iteration is

$$x_j^{k+1} = (1 - m_n^{-1}s_{nj})x_j^k + m_n^{-1}x_j^k \sum_{i \in B_n} A_{ij}\Big(\frac{b_i}{(Ax^k)_i}\Big). \tag{13.10}$$

Actually, all we need is that the parameter m_n be chosen so that $s_{nj} \leq m_n$; with the choice of $m_n = \max\{s_{nj}|j = 1, 2, ..., J\}$ the algorithm is called the *rescaled block-iterative* EMML (RBI-EMML) [54]. Notice that when $s_{nj} = t_n r_j$, the first term vanishes, since $m_n^{-1}s_{nj} = 1$, and the RBI-EMML becomes the OSEM.

13.4.4 The EMART Algorithm

When we apply the formula for the RBI-EMML to the case of $N = I$, we obtain the analog of the MART that we have been seeking. It has the iterative step

$$x_j^{k+1} = (1 - m_i^{-1}A_{ij})x_j^k + m_i^{-1}A_{ij}\Big(x_j^k \frac{b_i}{(Ax^k)_i}\Big). \tag{13.11}$$

13.5 KL Projections

The term $x_j^k\Big(\frac{b_i}{(Ax^k)_i}\Big)$ shows up in all the algorithms we have considered so far in this chapter. It is reasonable to ask if this term has any significance.

The ART and Cimmino algorithms involve the orthogonal projections onto the hyperplanes determined by each of the equations in the system. Now we are considering non-negative systems of linear equations, so it makes sense to define

$$H_i = \{x \geq 0 | (Ax)_i = b_i\}.$$

When we try to calculate the KL projection of a vector $z \geq 0$ onto H_i, that is, when we try to find the member of H_i that minimizes $KL(x, z)$, we find that we cannot solve for x in closed form. However, suppose that we calculate the x in H_i that minimizes the distance

$$\sum_{j=1}^{J} A_{ij} KL(x_j, z_j),$$

the *weighted KL projection* of z onto H_i. We find that the solution is

$$x_j = z_j\Big(\frac{b_i}{(Az)_i}\Big).$$

Therefore, the term $x_j^k\Big(\frac{b_i}{(Ax^k)_i}\Big)$ is the vector in H_i that minimizes

$$\sum_{j=1}^{J} A_{ij} KL(x_j, x_j^k).$$

All the algorithms we have considered in this chapter rely on the weighted KL projection of the current vector onto H_i.

13.6 Some Open Questions

We know that the RBI-SMART algorithms converge to the non-negative solution of $Ax = b$ for which $KL(x, x^0)$ is minimized, for any choice of blocks, whenever $Ax = b$ has non-negative solutions. We know that the RBI-EMML algorithms converge to a non-negative solution of $Ax = b$, whenever $Ax = b$ has non-negative solutions. We do not know if the solution obtained depends on the blocks chosen, and we do not know which non-negative solution the algorithms give us, even in the case of the original EMML algorithm.

Chapter 14

The Split Feasibility Problem

14.1 Chapter Summary

The *split feasibility problem* (SFP) [81] is to find $c \in C$ with $Ac \in Q$, if such points exist, where A is a real I by J matrix and C and Q are nonempty, closed convex sets in $\mathbb{R}^J$ and $\mathbb{R}^I$, respectively. In this chapter we discuss the CQ algorithm for solving the SFP, as well as recent extensions and applications to radiation therapy.

14.2 The CQ Algorithm

In [62] the CQ algorithm for solving the SFP was presented, for the real case. It has the iterative step

$$x^{k+1} = P_C(x^k - \gamma A^T(I - P_Q)Ax^k), \tag{14.1}$$

where I is the identity operator and $\gamma \in (0, 2/\rho(A^TA))$, for $\rho(A^TA)$ the spectral radius of the matrix A^TA, which is also its largest eigenvalue. The CQ algorithm can be extended to the complex case, in which the matrix A

has complex entries, and the sets C and Q are in $\mathbb{C}^J$ and $\mathbb{C}^I$, respectively. The iterative step of the extended CQ algorithm is then

$$x^{k+1} = P_C(x^k - \gamma A^\dagger(I - P_Q)Ax^k). \tag{14.2}$$

The CQ algorithm converges to a solution of the SFP, for any starting vector x^0, whenever the SFP has solutions. When the SFP has no solutions, the CQ algorithm converges to a minimizer of the function

$$f(x) = \frac{1}{2}||P_Q Ax - Ax||_2^2$$

over the set C, provided such constrained minimizers exist [63]. The CQ algorithm employs the relaxation parameter γ in the interval $(0, 2/L)$, where L is the largest eigenvalue of the matrix $A^T A$. Choosing the best relaxation parameter in any algorithm is a nontrivial procedure. Generally speaking, we want to select γ near to $1/L$. If A is normalized so that each row has length one, then the spectral radius of $A^T A$ does not exceed the maximum number of nonzero elements in any column of A. A similar upper bound on $\rho(A^T A)$ can be obtained for non-normalized, ϵ-sparse A.

14.3 Particular Cases of the CQ Algorithm

It is easy to find important examples of the SFP: if $C \subseteq \mathbb{R}^J$ and $Q = \{b\}$ then solving the SFP amounts to solving the linear system of equations $Ax = b$; if C is a proper subset of $\mathbb{R}^J$, such as the nonnegative cone, then we seek solutions of $Ax = b$ that lie within C, if there are any. Generally, we cannot solve the SFP in closed form and iterative methods are needed.

A number of well known iterative algorithms, such as the Landweber [187] and projected Landweber methods (see [18]), are particular cases of the CQ algorithm.

14.3.1 The Landweber algorithm

With x^0 arbitrary and $k = 0, 1, ...$ let

$$x^{k+1} = x^k + \gamma A^T(b - Ax^k). \tag{14.3}$$

This is the Landweber algorithm.

14.3.2 The Projected Landweber Algorithm

For a general nonempty closed convex C, x^0 arbitrary, and $k = 0, 1, ...$, the projected Landweber (PLW) algorithm for finding a solution of $Ax = b$ in C has the iterative step

$$x^{k+1} = P_C(x^k + \gamma A^T(b - Ax^k)). \tag{14.4}$$

14.3.3 Convergence of the Landweber Algorithms

From the convergence theorem for the CQ algorithm it follows that the Landweber algorithm converges to a solution of $Ax = b$ and the projected Landweber algorithm converges to a solution of $Ax = b$ in C, whenever such solutions exist. When there are no solutions of the desired type, the Landweber algorithm converges to a least squares approximate solution of $Ax = b$, while the projected Landweber algorithm will converge to a minimizer, over the set C, of the function $||b - Ax||_2$, whenever such a minimizer exists.

14.3.4 The Simultaneous ART (SART)

Another example of the CQ algorithm is the *simultaneous algebraic reconstruction technique* (SART) [5] for solving $Ax = b$, for nonnegative matrix A. Let A be an I by J matrix with nonnegative entries. Let $A_{i+} > 0$ be the sum of the entries in the ith row of A and $A_{+j} > 0$ be the sum of the entries in the jth column of A. Consider the (possibly inconsistent) system $Ax = b$. The SART algorithm has the following iterative step:

$$x_j^{k+1} = x_j^k + \frac{1}{A_{+j}} \sum\nolimits_{i=1}^{I} A_{ij}(b_i - (Ax^k)_i)/A_{i+}.$$

We make the following changes of variables:

$$B_{ij} = A_{ij}/(A_{i+})^{1/2}(A_{+j})^{1/2},$$

$$z_j = x_j(A_{+j})^{1/2},$$

and

$$c_i = b_i/(A_{i+})^{1/2}.$$

Then the SART iterative step can be written as

$$z^{k+1} = z^k + B^T(c - Bz^k).$$

This is a particular case of the Landweber algorithm, with $\gamma = 1$. The convergence of SART follows from that of the CQ algorithm, once we know

that the largest eigenvalue of B^TB is less than two; in fact, we show that it is one [62].

If B^TB had an eigenvalue greater than one and some of the entries of A are zero, then, replacing these zero entries with very small positive entries, we could obtain a new A whose associated B^TB also had an eigenvalue greater than one. Therefore, we assume, without loss of generality, that A has all positive entries. Since the new B^TB also has only positive entries, this matrix is irreducible and the Perron-Frobenius Theorem applies. We shall use this to complete the proof.

Let $u = (u_1, ..., u_J)^T$ with $u_j = (A_{+j})^{1/2}$ and $v = (v_1, ..., v_I)^T$, with $v_i = (A_{i+})^{1/2}$. Then we have $Bu = v$ and $B^Tv = u$; that is, u is an eigenvector of B^TB with associated eigenvalue equal to one, and all the entries of u are positive, by assumption. The Perron-Frobenius theorem applies and tells us that the eigenvector associated with the largest eigenvalue has all positive entries. Since the matrix B^TB is symmetric its eigenvectors are orthogonal; therefore u itself must be an eigenvector associated with the largest eigenvalue of B^TB. The convergence of SART follows.

14.3.5 Application of the CQ Algorithm in Dynamic ET

To illustrate how an image reconstruction problem can be formulated as a SFP, we consider briefly *emission computed tomography* (ET) image reconstruction. The objective in ET is to reconstruct the internal spatial distribution of intensity of a radionuclide from counts of photons detected outside the patient. In static ET the intensity distribution is assumed constant over the scanning time. Our data are photon counts at the detectors, forming the positive vector b and we have a matrix A of detection probabilities; our model is $Ax = b$, for x a nonnegative vector. We could then take $Q = \{b\}$ and $C = \mathbb{R}^N_+$, the nonnegative cone in $\mathbb{R}^N$.

In *dynamic* ET [129] the intensity levels at each voxel may vary with time. The observation time is subdivided into, say, T intervals and one static image, call it x^t, is associated with the time interval denoted by t, for $t = 1, ..., T$. The vector x is the concatenation of these T image vectors x^t. The discrete time interval at which each data value is collected is also recorded and the problem is to reconstruct this succession of images.

Because the data associated with a single time interval is insufficient, by itself, to generate a useful image, one often uses prior information concerning the time history at each fixed voxel to devise a model of the behavior of the intensity levels at each voxel, as functions of time. One may, for example, assume that the radionuclide intensities at a fixed voxel are increasing with time, or are concave (or convex) with time. The problem then is to find $x \geq 0$ with $Ax = b$ and $Dx \geq 0$, where D is a matrix chosen to describe this additional prior information. For example, we may wish to require that, for each fixed voxel, the intensity is an increasing function of

(discrete) time; then we want

$$x_j^{t+1} - x_j^t \geq 0,$$

for each t and each voxel index j. Or, we may wish to require that the intensity at each voxel describes a concave function of time, in which case nonnegative second differences would be imposed:

$$(x_j^{t+1} - x_j^t) - (x_j^{t+2} - x_j^{t+1}) \geq 0.$$

In either case, the matrix D can be selected to include the left sides of these inequalities, while the set Q can include the nonnegative cone as one factor.

14.3.6 More on the CQ Algorithm

One of the obvious drawbacks to the use of the CQ algorithm is that we would need the projections P_C and P_Q to be easily calculated. Several authors have offered remedies for that problem, using approximations of the convex sets by the intersection of hyperplanes and orthogonal projections onto those hyperplanes [268].

14.3.7 Convex Feasibility and IMRT

The CQ algorithm [62, 63] is an iterative algorithm for solving the split feasibility problem. Because it is particularly simple to implement in many cases, it has become the focus of recent work in intensity modulated radiation therapy (IMRT). In [84] Censor *et al.* extend the CQ algorithm to solve what they call the *multiple-set split feasibility problem* (MSSFP) . In the sequel [82] it is shown that the constraints in IMRT can be modeled as inclusion in convex sets and the extended CQ algorithm is used to determine dose intensities for IMRT that satisfy both dose constraints and radiation-source constraints.

14.4 Applications of the PLW Algorithm

Suppose that G is an arbitrary I by J matrix, and that $D \subseteq \mathbb{C}^J$ is a closed, non-empty convex set. We can use the PLW algorithm to minimize $\|Gw\|_2$ over $w \in D$: the iterative step is

$$w^{k+1} = P_D(w^k - \gamma G^\dagger G w^k), \tag{14.5}$$

for $0 < \gamma < \frac{2}{\rho(G^\dagger G)}$. The sequence $\{w^k\}$ converges to a minimizer, over $w \in D$, of $\|Gw\|_2$, whenever such minimizers exist.

Suppose now that A is an M by N matrix, and B an M by K matrix. Suppose also that $C \subseteq \mathbb{C}^N$, and $Q \subseteq \mathbb{C}^M$ are closed, non-empty convex sets. We want to find $x \in C$ and $y \in Q$ with $Ax = By$. Failing that, we want to minimize $\|Ax - By\|_2$ over $x \in C$ and $y \in Q$.

Let $G = \begin{bmatrix} A & -B \end{bmatrix}$ and $w = \begin{bmatrix} x \\ y \end{bmatrix}$ in $\mathbb{C}^{N+K}$. Then $Gw = Ax - By$. We apply the iteration in Equation (14.5) to minimize $\|Gw\|_2$ over $w \in D = C \times Q$, or, equivalently, to minimize $\|Ax - By\|_2$ over $x \in C$ and $y \in Q$.

We have

$$G^\dagger G = \begin{bmatrix} A^\dagger A & -A^\dagger B \\ -B^\dagger A & B^\dagger B \end{bmatrix},$$

so that the iteration in Equation (14.5) becomes

$$x^{k+1} = P_C(x^k - \gamma A^\dagger(Ax^k - By^k)), \tag{14.6}$$

and

$$y^{k+1} = P_Q(y^k + \gamma B^\dagger(Ax^k - By^k)). \tag{14.7}$$

Chapter 15

Conjugate-Direction Methods

15.1 Chapter Summary

Finding the least-squares solution of a possibly inconsistent system of linear equations $Ax = b$ is equivalent to minimizing the quadratic function $f(x) = \frac{1}{2}||Ax - b||_2^2$ and so can be viewed within the framework of optimization. Iterative optimization methods can then be used to provide, or at least suggest, algorithms for obtaining the least-squares solution. The *conjugate gradient method* is one such method.

15.2 Iterative Minimization

Iterative methods for minimizing a real-valued function $f(x)$ over the vector variable x usually take the following form: having obtained x^{k-1}, a new direction vector d^k is selected, an appropriate scalar $\alpha_k > 0$ is determined and the next member of the iterative sequence is given by

$$x^k = x^{k-1} + \alpha_k d^k. \tag{15.1}$$

Ideally, one would choose the α_k to be the value of α for which the function $f(x^{k-1}+\alpha d^k)$ is minimized. It is assumed that the direction d^k is a *descent direction*; that is, for small positive α the function $f(x^{k-1}+\alpha d^k)$ is strictly decreasing. Finding the optimal value of α at each step of the iteration is difficult, if not impossible, in most cases, and approximate methods, using line searches, are commonly used.

Ex. 15.1 *Differentiate the function $f(x^{k-1}+\alpha d^k)$ with respect to the variable α to show that, when $\alpha = \alpha_k$ is optimal, then*

$$\nabla f(x^k) \cdot d^k = 0. \tag{15.2}$$

Since the gradient $\nabla f(x^k)$ is orthogonal to the previous direction vector d^k and also because $-\nabla f(x)$ is the direction of greatest decrease of $f(x)$, the choice of $d^{k+1} = -\nabla f(x^k)$ as the next direction vector is a reasonable one. With this choice we obtain Cauchy's *steepest descent method* [198]:

$$x^{k+1} = x^k - \alpha_{k+1}\nabla f(x^k).$$

The steepest descent method need not converge in general and even when it does, it can do so slowly, suggesting that there may be better choices for the direction vectors. For example, the Newton-Raphson method [209] employs the following iteration:

$$x^{k+1} = x^k - \nabla^2 f(x^k)^{-1}\nabla f(x^k),$$

where $\nabla^2 f(x)$ is the Hessian matrix for $f(x)$ at x. To investigate further the issues associated with the selection of the direction vectors, we consider the more tractable special case of quadratic optimization.

15.3 Quadratic Optimization

Let A be an arbitrary real I by J matrix. The linear system of equations $Ax = b$ need not have any solutions, and we may wish to find a least-squares solution $x = \hat{x}$ that minimizes

$$f(x) = \frac{1}{2}||b - Ax||_2^2. \tag{15.3}$$

The vector b can be written

$$b = A\hat{x} + \hat{w},$$

where $A^T\hat{w} = 0$ and a least squares solution is an exact solution of the linear system $Qx = c$, with $Q = A^TA$ and $c = A^Tb$. We shall assume that Q is invertible and there is a unique least squares solution; this is the typical case.

We consider now the iterative scheme described by Equation (15.1) for $f(x)$ as in Equation (15.3). For this $f(x)$ the gradient becomes

$$\nabla f(x) = Qx - c.$$

The optimal α_k for the iteration can be obtained in closed form.

Ex. 15.2 *Show that the optimal α_k is*

$$\alpha_k = \frac{r^k \cdot d^k}{d^k \cdot Qd^k}, \tag{15.4}$$

where $r^k = c - Qx^{k-1}$.

Ex. 15.3 *Let $||x||_Q^2 = x \cdot Qx$ denote the square of the Q-norm of x. Show that*

$$||\hat{x} - x^{k-1}||_Q^2 - ||\hat{x} - x^k||_Q^2 = (r^k \cdot d^k)^2/d^k \cdot Qd^k \geq 0$$

for any direction vectors d^k.

If the sequence of direction vectors $\{d^k\}$ is completely general, the iterative sequence need not converge. However, if the set of direction vectors is finite and spans $\mathbb{R}^J$ and we employ them cyclically, convergence follows.

Theorem 15.1 *Let $\{d^1, ..., d^J\}$ be any finite set whose span is all of $\mathbb{R}^J$. Let α_k be chosen according to Equation (15.4). Then, for $k = 1, 2, ...$, $j = k(\mathrm{mod}\, J)$, and any x^0, the sequence defined by*

$$x^k = x^{k-1} + \alpha_k d^j$$

converges to the least squares solution.

Proof: The sequence $\{||\hat{x} - x^k||_Q^2\}$ is decreasing and, therefore, the sequence $\{(r^k \cdot d^k)^2/d^k \cdot Qd^k$ must converge to zero. Therefore, the vectors x^k are bounded, and for each $j = 1, ..., J$, the subsequences $\{x^{mJ+j}, m = 0, 1, ...\}$ have cluster points, say $x^{*,j}$ with

$$x^{*,j} = x^{*,j-1} + \frac{(c - Qx^{*,j-1}) \cdot d^j}{d^j \cdot Qd^j} d^j.$$

Since

$$r^{mJ+j} \cdot d^j \to 0,$$

it follows that, for each $j = 1, ..., J$,

$$(c - Qx^{*,j}) \cdot d^j = 0.$$

Therefore,

$$x^{*,1} = ... = x^{*,J} = x^*$$

with $Qx^* = c$. Consequently, x^* is the least squares solution and the sequence $\{||x^* - x^k||_Q\}$ is decreasing. But a subsequence converges to zero; therefore, $\{||x^* - x^k||_Q\} \to 0$. This completes the proof. ∎

In the quadratic case the steepest descent iteration has the form

$$x^k = x^{k-1} + \frac{r^k \cdot r^k}{r^k \cdot Qr^k} r^k.$$

We have the following result.

Theorem 15.2 *The steepest descent method converges to the least-squares solution.*

Proof: As in the proof of the previous theorem, we have

$$||\hat{x} - x^{k-1}||_Q^2 - ||\hat{x} - x^k||_Q^2 = (r^k \cdot d^k)^2 / d^k \cdot Qd^k \geq 0,$$

where now the direction vectors are $d^k = r^k$. So, the sequence $\{||\hat{x} - x^k||_Q^2\}$ is decreasing, and therefore the sequence $\{(r^k \cdot r^k)^2 / r^k \cdot Qr^k\}$ must converge to zero. The sequence $\{x^k\}$ is bounded; let x^* be a cluster point. It follows that $c - Qx^* = 0$, so that x^* is the least-squares solution $\hat{x}$. The rest of the proof follows as in the proof of the previous theorem. ∎

There is an interesting corollary to the theorem that pertains to a modified version of the ART algorithm. For $k = 1, 2, ...$ and $i = k(\mathrm{mod}\, M)$ and with the rows of A normalized to have length one, the ART iterative step is

$$x^k = x^{k-1} + (b_i - (Ax^{k-1})_i)a^i,$$

where a^i is the ith column of A^T. When $Ax = b$ has no solutions, the ART algorithm does not converge to the least-squares solution; rather, it exhibits subsequential convergence to a limit cycle. However, using the previous theorem, we can show that the following modification of the ART, which we shall call the *least squares ART* (LS-ART), converges to the least-squares solution for every x^0:

$$x^k = x^{k-1} + \frac{r^k \cdot a^i}{a^i \cdot Qa^i} a^i.$$

15.4 Conjugate Bases for $\mathbb{R}^J$

If the set $\{v^1, ..., v^J\}$ is a basis for $\mathbb{R}^J$, then any vector x in $\mathbb{R}^J$ can be expressed as a linear combination of the basis vectors; that is, there are real numbers $a_1, ..., a_J$ for which

$$x = a_1 v^1 + a_2 v^2 + ... + a_J v^J.$$

For each x the coefficients a_j are unique. To determine the a_j we write

$$x \cdot v^m = a_1 v^1 \cdot v^m + a_2 v^2 \cdot v^m + ... + a_J v^J \cdot v^m,$$

for $m = 1, ..., J$. Having calculated the quantities $x \cdot v^m$ and $v^j \cdot v^m$, we solve the resulting system of linear equations for the a_j.

If, instead of an arbitrary basis $\{v^1, ..., v^J\}$, we use an orthogonal basis $\{u^1, ..., u^J\}$, that is, $u^j \cdot u^m = 0$, unless $j = m$, then the system of linear equations is trivial to solve. The solution is $a_j = x \cdot u^j / u^j \cdot u^j$, for each j. Of course, we still need to compute the quantities $x \cdot u^j$.

The least-squares solution of the linear system of equations $Ax = b$ is

$$\hat{x} = (A^T A)^{-1} A^T b = Q^{-1} c.$$

To express $\hat{x}$ as a linear combination of the members of an orthogonal basis $\{u^1, ..., u^J\}$ we need the quantities $\hat{x} \cdot u^j$, which usually means that we need to know $\hat{x}$ first. For a special kind of basis, a *Q-conjugate basis*, knowing $\hat{x}$ ahead of time is not necessary; we need only know Q and c. Therefore, we can use such a basis to find $\hat{x}$. This is the essence of the *conjugate gradient method* (CGM), in which we calculate a conjugate basis and, in the process, determine $\hat{x}$.

15.4.1 Conjugate Directions

From Equation (15.2) we have

$$(c - Qx^k) \cdot d^k = 0,$$

which can be expressed as

$$(\hat{x} - x^k) \cdot Qd^k = (\hat{x} - x^k)^T Qd^k = 0.$$

Two vectors x and y are said to be *Q-orthogonal* (or *Q-conjugate*, or just *conjugate*) if $x \cdot Qy = 0$. So, the least-squares solution that we seek lies in a direction from x^k that is Q-orthogonal to d^k. This suggests that we can do better than steepest descent if we take the next direction to be Q-orthogonal to the previous one, rather than just orthogonal. This leads us to *conjugate direction methods*.

Ex. 15.4 *Say that the set* $\{p^1, ..., p^n\}$ *is a conjugate set for* $\mathbb{R}^J$ *if* $p^i \cdot Qp^j = 0$ *for* $i \neq j$. *Prove that a conjugate set that does not contain zero is linearly independent. Show that if* $p^n \neq 0$ *for* $n = 1, ..., J$, *then the least-squares vector* $\hat{x}$ *can be written as*

$$\hat{x} = a_1 p^1 + ... + a_J p^J,$$

with $a_j = c \cdot p^j / p^j \cdot Qp^j$ *for each* j. *Hint: use the* Q*-inner product* $\langle x, y \rangle_Q = x \cdot Qy$.

Therefore, once we have a conjugate basis, computing the least squares solution is trivial. Generating a conjugate basis can obviously be done using the standard Gram-Schmidt approach.

15.4.2 The Gram-Schmidt Method

Let $\{v^1, ..., v^J\}$ be a basis for the space $\mathbb{R}^J$. The Gram-Schmidt method uses the v^j to create an orthogonal basis $\{u^1, ..., u^J\}$ for $\mathbb{R}^J$. Begin by taking $u^1 = v^1$. For $j = 2, ..., J$, let

$$u^j = v^j - \frac{u^1 \cdot v^j}{u^1 \cdot u^1} u^1 - ... - \frac{u^{j-1} \cdot v^j}{u^{j-1} \cdot u^{j-1}} u^{j-1}.$$

To apply this approach to obtain a conjugate basis, we would simply replace the dot products $u^k \cdot v^j$ and $u^k \cdot u^k$ with the Q-inner products, that is,

$$p^j = v^j - \frac{p^1 \cdot Qv^j}{p^1 \cdot Qp^1} p^1 - ... - \frac{p^{j-1} \cdot Qv^j}{p^{j-1} \cdot Qp^{j-1}} p^{j-1}. \tag{15.5}$$

Even though the Q-inner products can always be written as $x \cdot Qy = Ax \cdot Ay$, so that we need not compute the matrix Q, calculating a conjugate basis using Gram-Schmidt is not practical for large J. There is a way out, fortunately.

If we take $p^1 = v^1$ and $v^j = Qp^{j-1}$, we have a much more efficient mechanism for generating a conjugate basis, namely a three-term recursion formula [198]. The set $\{p^1, Qp^1, ..., Qp^{J-1}\}$ need not be a linearly independent set, in general, but, if our goal is to find $\hat{x}$, and not really to calculate a full conjugate basis, this does not matter, as we shall see.

Theorem 15.3 *Let* $p^1 \neq 0$ *be arbitrary. Let* p^2 *be given by*

$$p^2 = Qp^1 - \frac{Qp^1 \cdot Qp^1}{p^1 \cdot Qp^1} p^1,$$

so that $p^2 \cdot Qp^1 = 0$. *Then, for* $n \geq 2$, *let* p^{n+1} *be given by*

$$p^{n+1} = Qp^n - \frac{Qp^n \cdot Qp^n}{p^n \cdot Qp^n} p^n - \frac{Qp^{n-1} \cdot Qp^n}{p^{n-1} \cdot Qp^{n-1}} p^{n-1}. \tag{15.6}$$

Then, the set $\{p^1, ..., p^J\}$ *is a conjugate set for* $\mathbb{R}^J$*. If* $p^n \neq 0$ *for each* n*, then the set is a conjugate basis for* $\mathbb{R}^J$*.*

Proof: We consider the induction step of the proof. Assume that $\{p^1, ..., p^n\}$ is a Q-orthogonal set of vectors; we then show that $\{p^1, ..., p^{n+1}\}$ is also, provided that $n \leq J - 1$. It is clear from Equation (15.6) that

$$p^{n+1} \cdot Qp^n = p^{n+1} \cdot Qp^{n-1} = 0.$$

For $j \leq n - 2$, we have

$$p^{n+1} \cdot Qp^j = p^j \cdot Qp^{n+1} = p^j \cdot Q^2 p^n - ap^j \cdot Qp^n - bp^j \cdot Qp^{n-1},$$

for constants a and b. The second and third terms on the right side are then zero because of the induction hypothesis. The first term is also zero since

$$p^j \cdot Q^2 p^n = (Qp^j) \cdot Qp^n = 0$$

because Qp^j is in the span of $\{p^1, ..., p^{j+1}\}$, and so is Q-orthogonal to p^n. ∎

The calculations in the three-term recursion formula Equation (15.6) also occur in the Gram-Schmidt approach in Equation (15.5); the point is that Equation (15.6) uses only the first three terms, in every case.

15.5 The Conjugate Gradient Method

The main idea in the *conjugate gradient method* (CGM) is to build the conjugate set as we calculate the least squares solution using the iterative algorithm

$$x^n = x^{n-1} + \alpha_n p^n. \tag{15.7}$$

The α_n is chosen so as to minimize $f(x^{n-1} + \alpha p^n)$ as a function of α, and so we have

$$\alpha_n = \frac{r^n \cdot p^n}{p^n \cdot Qp^n}, \tag{15.8}$$

where $r^n = c - Qx^{n-1}$.

Ex. 15.5 *Show that*

$$r^{n+1} = r^n - \alpha_n Qp^n, \tag{15.9}$$

so Qp^n *is in the span of* r^{n+1} *and* r^n*.*

Since the function $f(x) = \frac{1}{2}||Ax - b||_2^2$ has for its gradient $\nabla f(x) = A^T(Ax - b) = Qx - c$, the residual vector $r^n = c - Qx^{n-1}$ is the direction of steepest descent from the point $x = x^{n-1}$. The CGM combines the use of the negative gradient directions from the steepest descent method with the use of a conjugate basis of directions, by using the r^n to construct the next direction p^n in such a way as to form a conjugate set $\{p^1, ..., p^J\}$.

As before, there is an efficient recursive formula that provides the next direction: let $p^1 = r^1 = (c - Qx^0)$ and for $j = 2, 3, ...$

$$p^j = r^j - \beta_{j-1}p^{j-1}, \tag{15.10}$$

with

$$\beta_{j-1} = \frac{r^j \cdot Qp^{j-1}}{p^{j-1} \cdot Qp^{j-1}}. \tag{15.11}$$

It follows from the definition of β_{j-1} that

$$p^j Qp^{j-1} = 0. \tag{15.12}$$

Since the α_n is the optimal choice and

$$r^{n+1} = -\nabla f(x^n),$$

we have, according to Equation (15.2),

$$r^{n+1} \cdot p^n = 0. \tag{15.13}$$

Ex. 15.6 *Prove that $r^n = 0$ whenever $p^n = 0$, in which case we have $c = Qx^{n-1}$, so that x^{n-1} is the least-squares solution.*

Ex. 15.7 *Show that $r^n \cdot p^n = r^n \cdot r^n$, so that*

$$\alpha_n = \frac{r^n \cdot r^n}{p^n \cdot Qp^n}. \tag{15.14}$$

In theory, the CGM converges to the least squares solution in finitely many steps, since we either reach $p^{n+1} = 0$ or $n + 1 = J$. With $x^0 = 0$ and

$$x^n = x^{n-1} + \alpha_n p^n, \tag{15.15}$$

for $n = 1, 2, ..., J$, we have $x^J = \hat{x}$, the least squares solution. In practice, the CGM can be employed as a fully iterative method by cycling back through the previously used directions.

An induction proof similar to the one used to prove Theorem 15.3 establishes that the set $\{p^1, ..., p^J\}$ is a conjugate set [198, 209]. In fact, we can say more.

Theorem 15.4 *For $n = 1, 2, ..., J$ and $j = 1, ..., n-1$ we have*

- **a)** $r^n \cdot r^j = 0$*;*
- **b)** $r^n \cdot p^j = 0$*; and*
- **c)** $p^n \cdot Qp^j = 0$.

The proof presented here through a series of exercises is based on that given in [209].

The proof uses induction on the number n. Throughout the following exercises assume that the statements in the theorem hold for some fixed n with $2 \leq n < J$ and for $j = 1, 2, ..., n-1$. We prove that they hold also for $n+1$ and $j = 1, 2, ..., n$.

Ex. 15.8 *Show that $p^n \cdot Qp^n = r^n \cdot Qp^n$, so that*

$$\alpha_n = \frac{r^n \cdot r^n}{r^n \cdot Qp^n}. \tag{15.16}$$

Hints: use Equation (15.10) and the induction assumption concerning **c)** *of the Theorem.*

Ex. 15.9 *Show that $r^{n+1} \cdot r^n = 0$. Hint: use Equations (15.16) and (15.9).*

Ex. 15.10 *Show that $r^{n+1} \cdot r^j = 0$, for $j = 1, ..., n-1$. Hints: write out r^{n+1} using Equation (15.9) and r^j using Equation (15.10), and use the induction hypotheses.*

Ex. 15.11 *Show that $r^{n+1} \cdot p^j = 0$, for $j = 1, ..., n$. Hints: use Equations (15.9) and (15.10) and induction assumptions* **b)** *and* **c)**.

Ex. 15.12 *Show that $p^{n+1} \cdot Qp^j = 0$, for $j = 1, ..., n-1$. Hints: use Equation (15.9), the previous exercise, and the induction assumptions.*

The final step in the proof is to show that $p^{n+1} \cdot Qp^n = 0$. But this follows immediately from Equation (15.12).

15.6 Krylov Subspaces

Another approach to deriving the conjugate gradient method is to use Krylov subspaces. If we select $x^0 = 0$ as our starting vector for the CGM, then $p^1 = r^1 = c$, and each p^{n+1} and x^{n+1} lie in the *Krylov subspace* $\mathcal{K}_n(Q, c)$, defined to be the span of the vectors $\{c, Qc, Q^2c, ..., Q^nc\}$.

For any x in $\mathbb{R}^J$, we have

$$\|x - \hat{x}\|_Q^2 = (x - \hat{x})^T Q(x - \hat{x}).$$

Minimizing $\|x - \hat{x}\|_Q^2$ over all x in $\mathcal{K}_n(Q, c)$ is equivalent to minimizing the same function over all x of the form $x = x^n + \alpha p^{n+1}$. This, in turn, is equivalent to minimizing

$$-2\alpha p^{n+1} \cdot r^{n+1} + \alpha^2 p^{n+1} \cdot Q p^{n+1},$$

over all α, which has for its solution the value $\alpha = \alpha_{n+1}$ used to calculate x^{n+1} in the CGM.

15.7 Convergence Issues

The convergence rate of the CGM depends on the condition number of the matrix Q, which is the ratio of its largest to its smallest eigenvalues. When the condition number is much greater than one convergence can be accelerated by *preconditioning* the matrix Q; this means replacing Q with $P^{-1/2}QP^{-1/2}$, for some positive-definite approximation P of Q (see [7]).

15.8 Extending the CGM

There are versions of the CGM for the minimization of nonquadratic functions. In the quadratic case the next conjugate direction p^{n+1} is built from the residual r^{n+1} and p^n. Since, in that case, $r^{n+1} = -\nabla f(x^n)$, this suggests that in the nonquadratic case we build p^{n+1} from $-\nabla f(x^n)$ and p^n. This leads to the Fletcher-Reeves method. Other similar algorithms, such as the Polak-Ribiere and the Hestenes-Stiefel methods, perform better on certain problems [209].

Chapter 16

Regularization

16.1 Chapter Summary

When we use an iterative algorithm, we want it to solve our problem. We also want the solution in a reasonable amount of time, and we want slight errors in the measurements to cause only slight perturbations in the calculated answer. We have already discussed the use of block-iterative methods to accelerate convergence. Now we turn to regularization as a means of reducing sensitivity to noise. Because a number of regularization methods can be derived using a Bayesian *maximum a posteriori* approach, regularization is sometimes treated under the heading of MAP methods; see, for example, [207, 225] and the discussion in [65]. Penalty functions are also used for regularization [132, 2, 3].

16.2 Where Does Sensitivity Come From?

We illustrate the sensitivity problem that can arise when the inconsistent system $Ax = b$ has more equations than unknowns. We take A to be I by J and we calculate the least-squares solution,

$$x_{LS} = (A^\dagger A)^{-1} A^\dagger b, \tag{16.1}$$

assuming that the J by J Hermitian, nonnegative-definite matrix $Q = (A^\dagger A)$ is invertible, and therefore positive-definite.

The matrix Q has the eigenvalue/eigenvector decomposition

$$Q = \lambda_1 u_1 u_1^\dagger + \cdots + \lambda_J u_J u_J^\dagger, \tag{16.2}$$

where the (necessarily positive) eigenvalues of Q are

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_J > 0, \tag{16.3}$$

and the vectors u_j are the corresponding orthonormal eigenvectors.

16.2.1 The Singular-Value Decomposition of A

The square roots $\sqrt{\lambda_j}$ are called the *singular values* of A. The *singular-value decomposition* (SVD) of A is similar to the eigenvalue/eigenvector decomposition of Q: we have

$$A = \sqrt{\lambda_1} u_1 v_1^\dagger + \cdots + \sqrt{\lambda_I} u_J v_J^\dagger, \tag{16.4}$$

where the v_j are particular eigenvectors of $AA^\dagger$. We see from the SVD that the quantities $\sqrt{\lambda_j}$ determine the relative importance of each term $u_j v_j^\dagger$.

The SVD is commonly used for compressing transmitted or stored images. In such cases, the rectangular matrix A is a discretized image. It is not uncommon for many of the lowest singular values of A to be nearly zero, and to be essentially insignificant in the reconstruction of A. Only those terms in the SVD for which the singular values are significant need to be transmitted or stored. The resulting images may be slightly blurred, but can be restored later, as needed.

When the matrix A is a finite model of a linear imaging system, there will necessarily be model error in the selection of A. Getting the dominant terms in the SVD nearly correct is much more important (and usually much easier) than getting the smaller ones correct. The problems arise when we try to invert the system, to solve $Ax = b$ for x.

16.2.2 The Inverse of $Q = A^\dagger A$

The inverse of Q can then be written

$$Q^{-1} = \lambda_1^{-1} u_1 u_1^\dagger + \cdots + \lambda_J^{-1} u_J u_J^\dagger, \tag{16.5}$$

so that, with $A^\dagger b = c$, we have

$$x_{LS} = \lambda_1^{-1}(u_1^\dagger c)u_1 + \cdots + \lambda_J^{-1}(u_J^\dagger c)u_J. \tag{16.6}$$

Because the eigenvectors are orthonormal, we can express $||A^\dagger b||_2^2 = ||c||_2^2$ as

$$||c||_2^2 = |u_1^\dagger c|^2 + \cdots + |u_J^\dagger c|^2, \tag{16.7}$$

and $||x_{LS}||_2^2$ as

$$||x_{LS}||_2^2 = \lambda_1^{-1}|u_1^\dagger c|^2 + \cdots + \lambda_J^{-1}|u_J^\dagger c|^2. \tag{16.8}$$

It is not uncommon for the eigenvalues of Q to be quite distinct, with some of them much larger than the others. When this is the case, we see that $||x_{LS}||_2$ can be much larger than $||c||_2$, because of the presence of the terms involving the reciprocals of the small eigenvalues. When the measurements b are essentially noise-free, we may have $|u_j^\dagger c|$ relatively small, for the indices near J, keeping the product $\lambda_j^{-1}|u_j^\dagger c|^2$ reasonable in size, but when the b becomes noisy, this may no longer be the case. The result is that those terms corresponding to the reciprocals of the smallest eigenvalues dominate the sum for x_{LS} and the norm of x_{LS} becomes quite large. The least-squares solution we have computed is essentially all noise and useless.

In our discussion of the ART, we saw that when we impose a non-negativity constraint on the solution, noise in the data can manifest itself in a different way. When A has more columns than rows, but $Ax = b$ has no non-negative solution, then, at least for those A having the *full-rank property*, the non-negatively constrained least-squares solution has at most $I - 1$ non-zero entries. This happens also with the EMML and SMART solutions. As with the ART, regularization can eliminate the problem.

16.2.3 Reducing the Sensitivity to Noise

As we just saw, the presence of small eigenvalues for Q and noise in b can cause $||x_{LS}||_2$ to be much larger than $||A^\dagger b||_2$, with the result that x_{LS} is useless. In this case, even though x_{LS} minimizes $||Ax - b||_2$, it does so by overfitting to the noisy b. To reduce the sensitivity to noise and thereby obtain a more useful approximate solution, we can *regularize* the problem.

It often happens in applications that, even when there is an exact solution of $Ax = b$, noise in the vector b makes such as exact solution undesirable; in such cases a *regularized solution* is usually used instead. Select

$\epsilon > 0$ and a vector p that is a prior estimate of the desired solution. Define

$$F_\epsilon(x) = (1-\epsilon)\|Ax-b\|_2^2 + \epsilon\|x-p\|_2^2. \tag{16.9}$$

Lemma 16.1 *The function F_ϵ always has a unique minimizer $\hat{x}_\epsilon$, given by*

$$\hat{x}_\epsilon = ((1-\epsilon)A^\dagger A + \epsilon I)^{-1}((1-\epsilon)A^\dagger b + \epsilon p); \tag{16.10}$$

this is a regularized solution of $Ax = b$. Here, p is a prior estimate of the desired solution. Note that the inverse above always exists.

Note that, if $p = 0$, then

$$\hat{x}_\epsilon = (A^\dagger A + \gamma^2 I)^{-1} A^\dagger b, \tag{16.11}$$

for $\gamma^2 = \frac{\epsilon}{1-\epsilon}$. The regularized solution has been obtained by modifying the formula for x_{LS}, replacing the inverse of the matrix $Q = A^\dagger A$ with the inverse of $Q + \gamma^2 I$. When ϵ is near zero, so is γ^2, and the matrices Q and $Q + \gamma^2 I$ are nearly equal. What is different is that the eigenvalues of $Q + \gamma^2 I$ are $\lambda_i + \gamma^2$, so that, when the eigenvalues are inverted, the reciprocal eigenvalues are no larger than $1/\gamma^2$, which prevents the norm of x_ϵ from being too large, and decreases the sensitivity to noise.

Lemma 16.2 *Let ϵ be in $(0,1)$, and let I be the identity matrix whose dimensions are understood from the context. Then*

$$((1-\epsilon)AA^\dagger + \epsilon I)^{-1} A = A((1-\epsilon)A^\dagger A + \epsilon I)^{-1}, \tag{16.12}$$

and, taking conjugate transposes,

$$A^\dagger((1-\epsilon)AA^\dagger + \epsilon I)^{-1} = ((1-\epsilon)A^\dagger A + \epsilon I)^{-1} A^\dagger. \tag{16.13}$$

Proof: Use the identity

$$A((1-\epsilon)A^\dagger A + \epsilon I) = ((1-\epsilon)AA^\dagger + \epsilon I)A. \tag{16.14}$$

∎

Lemma 16.3 *Any vector p in $\mathbb{R}^J$ can be written as $p = A^\dagger q + r$, where $Ar = 0$.*

What happens to $\hat{x}_\epsilon$ as ϵ goes to zero? This will depend on which case we are in:

Case 1: $J \leq I$, and we assume that $A^\dagger A$ is invertible; or

Case 2: $J > I$, and we assume that $AA^\dagger$ is invertible.

Lemma 16.4 *In Case 1, taking limits as $\epsilon \to 0$ on both sides of the expression for $\hat{x}_\epsilon$ gives $\hat{x}_\epsilon \to (A^\dagger A)^{-1}A^\dagger b$, the least squares solution of $Ax = b$.*

We consider Case 2 now. Write $p = A^\dagger q + r$, with $Ar = 0$. Then

$$\hat{x}_\epsilon = A^\dagger((1-\epsilon)AA^\dagger + \epsilon I)^{-1}((1-\epsilon)b+$$

$$\epsilon q) + ((1-\epsilon)A^\dagger A + \epsilon I)^{-1}(\epsilon r). \tag{16.15}$$

Lemma 16.5 (a) *We have*

$$((1-\epsilon)A^\dagger A + \epsilon I)^{-1}(\epsilon r) = r, \tag{16.16}$$

for all $\epsilon \in (0,1)$. **(b)** *Taking the limit of $\hat{x}_\epsilon$, as $\epsilon \to 0$, we get $\hat{x}_\epsilon \to A^\dagger(AA^\dagger)^{-1}b + r$. This is the solution of $Ax = b$ closest to p.*

Proof: For part (a) let

$$t_\epsilon = ((1-\epsilon)A^\dagger A + \epsilon I)^{-1}(\epsilon r). \tag{16.17}$$

Then, multiplying by A gives

$$At_\epsilon = A((1-\epsilon)A^\dagger A + \epsilon I)^{-1}(\epsilon r). \tag{16.18}$$

Now show that $At_\epsilon = 0$. For part (b) draw a diagram for the case of one equation in two unknowns. ∎

16.3 Iterative Regularization

It is often the case that the entries of the vector b in the system $Ax = b$ come from measurements, so are usually noisy. If the entries of b are noisy but the system $Ax = b$ remains consistent (which can easily happen in the under-determined case, with $J > I$), the ART begun at $x^0 = 0$ converges to the solution having minimum norm, but this norm can be quite large. The resulting solution is probably useless. Instead of solving $Ax = b$, we *regularize* by minimizing, for example, the function $F_\epsilon(x)$ given in Equation (16.9). For the case of $p = 0$, the solution to this problem is the vector $\hat{x}_\epsilon$ in Equation (16.11). However, we do not want to calculate $A^\dagger A + \gamma^2 I$, in order to solve

$$(A^\dagger A + \gamma^2 I)x = A^\dagger b, \tag{16.19}$$

when the matrix A is large. Fortunately, there are ways to find $\hat{x}_\epsilon$, using only the matrix A. We later we shall see how this might be accomplished using the ART; now we show how the Landweber algorithm can be used to calculate this regularized solution.

16.3.1 Regularizing Landweber's Algorithm

Our goal is to minimize the function in Equation (16.9), with $p = 0$. Notice that this is equivalent to minimizing the function

$$F(x) = ||Bx - c||_2^2, \tag{16.20}$$

for

$$B = \begin{bmatrix} A \\ \gamma I \end{bmatrix}, \tag{16.21}$$

and

$$c = \begin{bmatrix} b \\ 0 \end{bmatrix}, \tag{16.22}$$

where 0 denotes a column vector with all entries equal to zero and $\gamma = \frac{\epsilon}{1-\epsilon}$. The Landweber iteration for the problem $Bx = c$ is

$$x^{k+1} = x^k + \alpha B^T(c - Bx^k), \tag{16.23}$$

for $0 < \alpha < 2/\rho(B^TB)$, where $\rho(B^TB)$ is the spectral radius of B^TB. Equation (16.23) can be written as

$$x^{k+1} = (1 - \alpha\gamma^2)x^k + \alpha A^T(b - Ax^k). \tag{16.24}$$

16.4 A Bayesian View of Reconstruction

The EMML iterative algorithm maximizes the likelihood function for the case in which the entries of the data vector $b = (b_1, ..., b_I)^T$ are assumed to be samples of independent Poisson random variables with mean values $(Ax)_i$; here, A is an I by J matrix with nonnegative entries and $x = (x_1, ..., x_J)^T$ is the vector of nonnegative parameters to be estimated. Equivalently, it minimizes the Kullback-Leibler distance $KL(b, Ax)$. This situation arises in single photon emission tomography, where the b_i are the number of photons counted at each detector i, x is the vectorized image to be reconstructed and its entries x_j are (proportional to) the radionuclide intensity levels at each voxel j. When the signal-to-noise ratio is low, which is almost always the case in medical applications, maximizing likelihood can lead to unacceptably noisy reconstructions, particularly when J is larger than I. One way to remedy this problem is simply to halt the EMML algorithm after a few iterations, to avoid over-fitting the x to the

noisy data. A more mathematically sophisticated remedy is to employ a penalized-likelihood or Bayesian approach and seek a maximum *a posteriori* (MAP) estimate of x.

In the Bayesian approach we view x as an instance of a random vector having a probability density function $f(x)$. Instead of maximizing the likelihood given the data, we now maximize the posterior likelihood, given both the data and the prior distribution for x. This is equivalent to minimizing

$$F(x) = KL(b, Ax) - \log f(x). \tag{16.25}$$

The EMML algorithm is an example of an optimization method based on alternating minimization of a function $H(x, z) > 0$ of two vector variables. The alternating minimization works this way: let x and z be vector variables and $H(x, z) > 0$. If we fix z and minimize $H(x, z)$ with respect to x, we find that the solution is $x = z$, the vector we fixed; that is,

$$H(x, z) \geq H(z, z) \tag{16.26}$$

always. If we fix x and minimize $H(x, z)$ with respect to z, we get something new; call it Tx. The EMML algorithm has the iterative step $x^{k+1} = Tx^k$.

Obviously, we can't use an arbitrary function H; it must be related to the function $KL(b, Ax)$ that we wish to minimize, and we must be able to obtain each intermediate optimizer in closed form. The clever step is to select $H(x, z)$ so that $H(x, x) = KL(b, Ax)$, for any x. Now see what we have so far:

$$KL(b, Ax^k) = H(x^k, x^k) \geq H(x^k, x^{k+1}) \tag{16.27}$$

$$\geq H(x^{k+1}, x^{k+1}) = KL(b, Ax^{k+1}). \tag{16.28}$$

That tells us that the algorithm makes $KL(b, Ax^k)$ decrease with each iteration. The proof doesn't stop here, but at least it is now plausible that the EMML iteration could minimize $KL(b, Ax)$.

The function $H(x, z)$ used in the EMML case is the KL distance

$$H(x, z) = KL(r(x), q(z)) = \sum_{i=1}^{I} \sum_{j=i}^{J} KL(r(x)_{ij}, q(z)_{ij}); \tag{16.29}$$

we define, for each nonnegative vector x for which $(Ax)_i = \sum_{j=1}^{J} A_{ij}x_j > 0$, the arrays $r(x) = \{r(x)_{ij}\}$ and $q(x) = \{q(x)_{ij}\}$ with entries

$$r(x)_{ij} = x_j A_{ij} \frac{b_i}{(Ax)_i} \tag{16.30}$$

and

$$q(x)_{ij} = x_j A_{ij}. \tag{16.31}$$

With $x = x^k$ fixed, we minimize with respect to z to obtain the next EMML iterate x^{k+1}. Having selected the prior pdf $f(x)$, we want an iterative algorithm to minimize the function $F(x)$ in Equation (16.25). It would be a great help if we could mimic the alternating minimization formulation and obtain x^{k+1} by minimizing

$$KL(r(x^k), q(z)) - \log f(z) \tag{16.32}$$

with respect to z. Unfortunately, to be able to express each new x^{k+1} in closed form, we need to choose $f(x)$ carefully.

16.5 The Gamma Prior Distribution for x

In [190] Lange *et al.* suggest viewing the entries x_j as samples of independent gamma-distributed random variables. A gamma-distributed random variable x takes positive values and has for its pdf the *gamma distribution* defined for positive x by

$$\gamma(x) = \frac{1}{\Gamma(\alpha)}(\frac{\alpha}{\beta})^{\alpha} x^{\alpha-1} e^{-\alpha x/\beta}, \tag{16.33}$$

where α and β are positive parameters and Γ denotes the gamma function. The mean of such a gamma-distributed random variable is then $\mu = \beta$ and the variance is $\sigma^2 = \beta^2/\alpha$.

Lemma 16.6 *If the entries z_j of z are viewed as independent and gamma-distributed with means μ_j and variances σ_j^2, then minimizing the function in line (16.32) with respect to z is equivalent to minimizing the function*

$$KL(r(x^k), q(z)) + \sum_{j=1}^{J} \delta_j KL(\gamma_j, z_j), \tag{16.34}$$

for

$$\delta_j = \frac{\mu_j}{\sigma_j^2}, \; \gamma_j = \frac{\mu_j^2 - \sigma_j^2}{\mu_j}, \tag{16.35}$$

under the assumption that the latter term is positive.

The resulting regularized EMML algorithm is the following:

Algorithm 16.1 (γ-prior Regularized EMML) *Let x^0 be an arbitrary positive vector. Then let*

$$x_j^{k+1} = \frac{\delta_j}{\delta_j + s_j}\gamma_j + \frac{1}{\delta_j + s_j} x_j^k \sum_{i=1}^{I} A_{ij} b_i / (Ax^k)_i, \tag{16.36}$$

where $s_j = \sum_{i=1}^{I} A_{ij}$.

We see from Equation (16.36) that the MAP iteration using the gamma priors generates a sequence of estimates each entry of which is a convex combination or weighted arithmetic mean of the result of one EMML step and the prior estimate γ_j. Convergence of the resulting iterative sequence is established by Lange, Bahn and Little in [190]; see also [51].

16.6 The One-Step-Late Alternative

It may well happen that we do not wish to use the gamma priors model and prefer some other $f(x)$. Because we will not be able to find a closed form expression for the z minimizing the function in line (16.32), we need some other way to proceed with the alternating minimization. Green [154] has offered the *one-step-late* (OSL) alternative.

When we try to minimize the function in line (16.32) by setting the gradient to zero we replace the variable z that occurs in the gradient of the term $-\log f(z)$ with x^k, the previously calculated iterate. Then, we can solve for z in closed form to obtain the new x^{k+1}. Unfortunately, negative entries can result and convergence is not guaranteed. There is a sizable literature on the use of MAP methods for this problem. In [60] an interior point algorithm (IPA) is presented that avoids the OSL issue. In [208] the IPA is used to regularize transmission tomographic images.

16.7 Regularizing the SMART

The SMART algorithm is not derived as a maximum likelihood method, so regularized versions do not take the form of MAP algorithms. Nevertheless, in the presence of noisy data, the SMART algorithm suffers from the same problem that afflicts the EMML, overfitting to noisy data resulting in an unacceptably noisy image. As we saw earlier, there is a close connection between the EMML and SMART algorithms. This suggests that a

regularization method for SMART can be developed along the lines of the MAP with gamma priors used for EMML. Since the SMART is obtained by minimizing the function $KL(q(z), r(x^k))$ with respect to z to obtain x^{k+1}, it seems reasonable to attempt to derive a regularized SMART iterative scheme by minimizing

$$KL(q(z), r(x^k)) + \sum_{j=1}^{J} \delta_j KL(z_j, \gamma_j), \tag{16.37}$$

as a function of z, for selected positive parameters δ_j and γ_j. This leads to the following algorithm:

Algorithm 16.2 (Regularized SMART) *Let x^0 be an arbitrary positive vector. Then let*

$$\log x_j^{k+1} = \frac{\delta_j}{\delta_j + s_j} \log \gamma_j + \frac{1}{\delta_j + s_j} x_j^k \sum_{i=1}^{I} A_{ij} \log[b_i/(Ax^k)_i]. \tag{16.38}$$

In [51] it was shown that this iterative sequence converges to a minimizer of the function

$$KL(Ax, y) + \sum_{j=1}^{J} \delta_j KL(x_j, \gamma_j). \tag{16.39}$$

It is useful to note that, although it may be possible to rederive this minimization problem within the framework of Bayesian MAP estimation by carefully selecting a prior pdf for the vector x, we have not done so. The MAP approach is a special case of regularization through the use of penalty functions. These penalty functions need not arise through a Bayesian formulation of the parameter-estimation problem.

16.8 De Pierro's Surrogate-Function Method

In [110] Alvaro De Pierro presents a modified EMML algorithm that includes regularization in the form of a penalty function. His objective is the same as ours was in the case of regularized SMART: to embed the penalty term in the alternating minimization framework in such a way as to make it possible to obtain the next iterate in closed form. Because his *surrogate function* method has been used subsequently by others to obtain penalized likelihood algorithms [91], we consider his approach in some detail.

Let x and z be vector variables and $H(x, z) > 0$. Mimicking the behavior

of the function $H(x, z)$ used in Equation (16.29), we require that if we fix z and minimize $H(x, z)$ with respect to x, the solution should be $x = z$, the vector we fixed; that is, $H(x, z) \geq H(z, z)$ always. If we fix x and minimize $H(x, z)$ with respect to z, we should get something new; call it Tx. As with the EMML, the algorithm will have the iterative step $x^{k+1} = Tx^k$.

Summarizing, we see that we need a function $H(x, z)$ with the following properties:

- (1) $H(x, z) \geq H(z, z)$ for all x and z;
- (2) $H(x, x)$ is the function $F(x)$ we wish to minimize; and
- (3) minimizing $H(x, z)$ with respect to z for fixed x is easy.

The function to be minimized is

$$F(x) = KL(b, Ax) + g(x), \tag{16.40}$$

where $g(x) \geq 0$ is some penalty function. De Pierro uses penalty functions $g(x)$ of the form

$$g(x) = \sum_{l=1}^{p} f_l(\langle s_l, x \rangle). \tag{16.41}$$

Let us define the matrix S to have for its lth row the vector s_l^T. Then $\langle s_l, x \rangle = (Sx)_l$, the lth entry of the vector Sx. Therefore,

$$g(x) = \sum_{l=1}^{p} f_l((Sx)_l). \tag{16.42}$$

Let $\lambda_{lj} > 0$ with $\sum_{j=1}^{J} \lambda_{lj} = 1$, for each l.

Assume that the functions f_l are convex. Therefore, for each l, we have

$$f_l((Sx)_l) = f_l(\sum_{j=1}^{J} S_{lj} x_j) = f_l(\sum_{j=1}^{J} \lambda_{lj}(S_{lj}/\lambda_{lj})x_j) \tag{16.43}$$

$$\leq \sum_{j=1}^{J} \lambda_{lj} f_l((S_{lj}/\lambda_{lj})x_j). \tag{16.44}$$

Therefore,

$$g(x) \leq \sum_{l=1}^{p} \sum_{j=1}^{J} \lambda_{lj} f_l((S_{lj}/\lambda_{lj})x_j). \tag{16.45}$$

So we have replaced $g(x)$ with a related function in which the x_j occur separately, rather than just in the combinations $(Sx)_l$. But we aren't quite done yet.

We would like to take for De Pierro's $H(x,z)$ the function used in the EMML algorithm, plus the function

$$\sum_{l=1}^{p}\sum_{j=1}^{J}\lambda_{lj}f_l((S_{lj}/\lambda_{lj})z_j). \tag{16.46}$$

But there is one slight problem: we need $H(z,z) = F(z)$, which we don't have yet.

De Pierro's clever trick is to replace $f_l((S_{lj}/\lambda_{lj})z_j)$ with

$$f_l\Big((S_{lj}/\lambda_{lj})z_j - (S_{lj}/\lambda_{lj})x_j\Big) + f_l((Sx)_l). \tag{16.47}$$

So, De Pierro's function $H(x,z)$ is the sum of the $H(x,z)$ used in the EMML case and the function

$$\sum_{l=1}^{p}\sum_{j=1}^{J}\lambda_{lj}f_l\Big((S_{lj}/\lambda_{lj})z_j - (S_{lj}/\lambda_{lj})x_j\Big) + \sum_{l=1}^{p}f_l((Sx)_l). \tag{16.48}$$

Now he has the three properties he needs. Once he has computed x^k, he minimizes $H(x^k,z)$ by taking the gradient and solving the equations for the correct $z = Tx^k = x^{k+1}$. For the choices of f_l he discusses, these intermediate calculations can either be done in closed form (the quadratic case) or with a simple Newton-Raphson iteration (the logcosh case).

16.9 Block-Iterative Regularization

We saw previously that it is possible to obtain a regularized least-squares solution $\hat{x}_\epsilon$, and thereby avoid the limit cycle, using only the matrix A and the ART algorithm. This prompts us to ask if it is possible to find regularized SMART solutions using block-iterative variants of SMART. Similarly, we wonder if it is possible to do the same for EMML.

Open Question: Can we use the MART to find the minimizer of the function

$$KL(Ax,b) + \epsilon KL(x,p)? \tag{16.49}$$

More generally, can we obtain the minimizer using RBI-SMART?

Open Question: Can we use the RBI-EMML methods to obtain the minimizer of the function

$$KL(b, Ax) + \epsilon KL(p, x)? \tag{16.50}$$

There have been various attempts to include regularization in block-iterative methods, to reduce noise sensitivity and avoid limit cycles; the paper by Ahn and Fessler [2] is a good source, as is [3]. Most of these approaches have been *ad hoc*, with little or no theoretical basis. Typically, they simply modify each iterative step by including an additional term that appears to be related to the regularizing penalty function. The case of the ART is instructive, however. In that case, we obtained the desired iterative algorithm by using an augmented set of variables, not simply by modifying each step of the original ART algorithm. How to do this for the MART and the other block-iterative algorithms is not obvious.

Recall that the RAMLA method in Equation (26.52) is similar to the RBI-EMML algorithm, but employs a sequence of decreasing relaxation parameters, which, if properly chosen, will cause the iterates to converge to the minimizer of $KL(b, Ax)$, thereby avoiding the limit cycle. In [112] De Pierro and Yamaguchi present a regularized version of RAMLA, but without guaranteed convergence.

Part IV

Applications

Chapter 17

Transmission Tomography I

17.1 Chapter Summary

Our topic is now transmission tomography. This chapter will provide a detailed description of how the data is gathered, the mathematical model of the scanning process, and the problem to be solved. In the next chapter we shall study the various mathematical techniques needed to solve this problem and the manner in which these techniques are applied. The man in Figure 17.1 is Allan Cormack, who won the Nobel Prize in 1979 for inventing the CAT scan.

17.2 X-ray Transmission Tomography

Although transmission tomography is not limited to scanning living beings, we shall concentrate here on the use of x-ray tomography in medical diagnosis and the issues that concern us in that application. The mathematical formulation will, of course, apply more generally.

In x-ray tomography, x-rays are transmitted through the body along many lines. In some, but not all, cases, the lines will all lie in the same plane. The strength of the x-rays upon entering the body is assumed known, and the strength upon leaving the body is measured. This data can then be used to estimate the amount of attenuation the x-ray encountered along that

line, which is taken to be the integral, along that line, of the attenuation function. On the basis of these line integrals, we estimate the attenuation function. This estimate is presented to the physician as one or more two-dimensional images.

17.3 The Exponential-Decay Model

As an x-ray beam passes through the body, it encounters various types of matter, such as soft tissue, bone, ligaments, air, each weakening the beam to a greater or lesser extent. If the intensity of the beam upon entry is I_{in} and I_{out} is its lower intensity after passing through the body, then

$$I_{out} = I_{in} e^{-\int_L f},$$

where $f = f(x, y) \geq 0$ is the *attenuation function* describing the two-dimensional distribution of matter within the slice of the body being scanned and $\int_L f$ is the integral of the function f over the line L along which the x-ray beam has passed. To see why this is the case, imagine the line L parameterized by the variable s and consider the intensity function $I(s)$ as a function of s. For small $\Delta s > 0$, the drop in intensity from the start to the end of the interval $[s, s + \Delta s]$ is approximately proportional to the intensity $I(s)$, to the attenuation $f(s)$ and to Δs, the length of the interval; that is,

$$I(s) - I(s + \Delta s) \approx f(s) I(s) \Delta s.$$

Dividing by Δs and letting Δs approach zero, we get

$$I'(s) = -f(s) I(s).$$

Ex. 17.1 *Show that the solution to this differential equation is*

$$I(s) = I(0) \exp(-\int_{u=0}^{u=s} f(u) du).$$

Hint: Use an integrating factor.

From knowledge of I_{in} and I_{out}, we can determine $\int_L f$. If we know $\int_L f$ for every line in the x, y-plane we can reconstruct the attenuation function f. In the real world we know line integrals only approximately and only for finitely many lines. The goal in x-ray transmission tomography is to estimate the attenuation function $f(x, y)$ in the slice, from finitely many noisy measurements of the line integrals. We usually have prior information about the values that $f(x, y)$ can take on. We also expect to find sharp boundaries separating regions where the function $f(x, y)$ varies only slightly. Therefore, we need algorithms capable of providing such images.

17.4 Difficulties to be Overcome

There are several problems associated with this model. X-ray beams are not exactly straight lines; the beams tend to spread out. The x-rays are not monochromatic, and their various frequency components are attenuated at different rates, resulting in *beam hardening*, that is, changes in the spectrum of the beam as it passes through the object. The beams consist of photons obeying statistical laws, so our algorithms probably should be based on these laws. How we choose the line segments is determined by the nature of the problem; in certain cases we are somewhat limited in our choice of these segments. Patients move; they breathe, their hearts beat, and, occasionally, they shift position during the scan. Compensating for these motions is an important, and difficult, aspect of the image reconstruction process. Finally, to be practical in a clinical setting, the processing that leads to the reconstructed image must be completed in a short time, usually around fifteen minutes. This time constraint is what motivates viewing the three-dimensional attenuation function in terms of its two-dimensional slices.

As we shall see, the Fourier transform and the associated theory of convolution filters play important roles in the reconstruction of transmission tomographic images.

The data we actually obtain at the detectors are counts of detected photons. These counts are not the line integrals; they are random quantities whose means, or expected values, are related to the line integrals. The Fourier inversion methods for solving the problem ignore its statistical aspects; in contrast, other methods, such as likelihood maximization, are based on a statistical model that involves Poisson-distributed emissions.

17.5 Reconstruction from Line Integrals

We turn now to the underlying problem of reconstructing attenuation functions from line-integral data.

17.5.1 The Radon Transform

Our goal is to reconstruct the function $f(x, y) \geq 0$ from line-integral data. Let θ be a fixed angle in the interval $[0, \pi)$. Form the t, s-axis system with the positive t-axis making the angle θ with the positive x-axis, as shown in Figure 17.2. Each point (x, y) in the original coordinate system

has coordinates (t, s) in the second system, where the t and s are given by

$$t = x\cos\theta + y\sin\theta,$$

and

$$s = -x\sin\theta + y\cos\theta.$$

If we have the new coordinates (t, s) of a point, the old coordinates are (x, y) given by

$$x = t\cos\theta - s\sin\theta,$$

and

$$y = t\sin\theta + s\cos\theta.$$

We can then write the function f as a function of the variables t and s. For each fixed value of t, we compute the integral

$$\int_L f(x,y)ds = \int f(t\cos\theta - s\sin\theta, t\sin\theta + s\cos\theta)ds$$

along the single line L corresponding to the fixed values of θ and t. We repeat this process for every value of t and then change the angle θ and repeat again. In this way we obtain the integrals of f over every line L in the plane. We denote by $r_f(\theta, t)$ the integral

$$r_f(\theta,t) = \int_L f(x,y)ds.$$

The function $r_f(\theta, t)$ is called the *Radon transform* of f.

17.5.2 The Central Slice Theorem

For fixed θ the function $r_f(\theta, t)$ is a function of the single real variable t; let $R_f(\theta, \omega)$ be its Fourier transform. Then

$$R_f(\theta,\omega) = \int r_f(\theta,t)e^{i\omega t}dt$$

$$= \int\int f(t\cos\theta - s\sin\theta, t\sin\theta + s\cos\theta)e^{i\omega t}dsdt$$

$$= \int\int f(x,y)e^{i\omega(x\cos\theta + y\sin\theta)}dxdy = F(\omega\cos\theta, \omega\sin\theta),$$

where $F(\omega\cos\theta, \omega\sin\theta)$ is the two-dimensional Fourier transform of the function $f(x, y)$, evaluated at the point $(\omega\cos\theta, \omega\sin\theta)$; this relationship is called the *Central Slice Theorem*. For fixed θ, as we change the value of ω, we obtain the values of the function F along the points of the line making the angle θ with the horizontal axis. As θ varies in $[0, \pi)$, we get all

the values of the function F. Once we have F, we can obtain f using the formula for the two-dimensional inverse Fourier transform. We conclude that we are able to determine f from its line integrals. As we shall see, inverting the Fourier transform can be implemented by combinations of frequency-domain filtering and back-projection.

FIGURE 17.1: Allan Cormack, who won the Nobel Prize for the CAT scan.

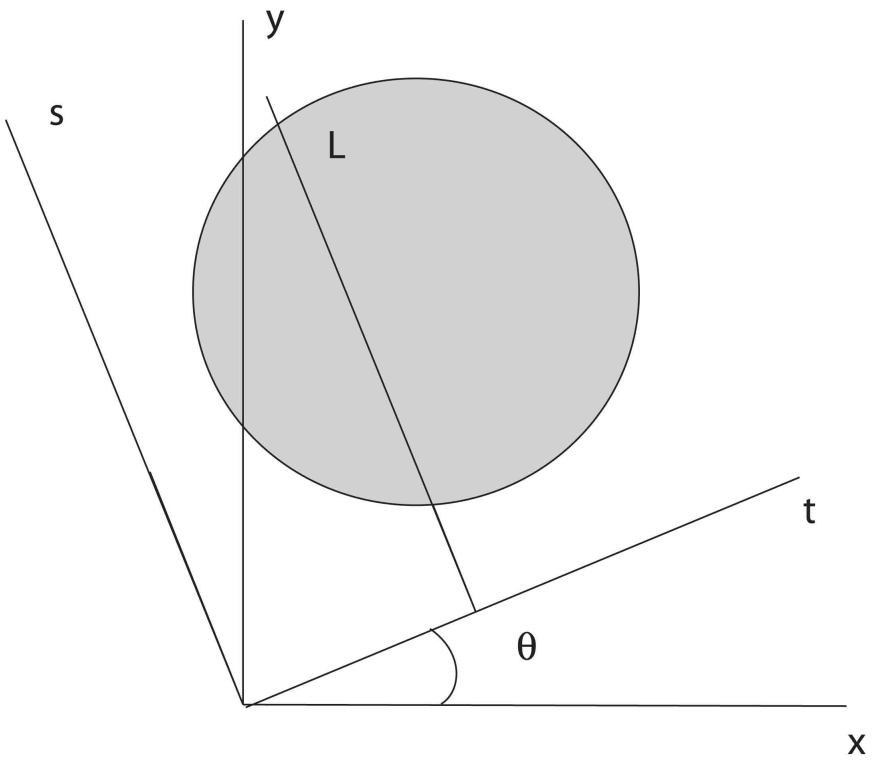

FIGURE 17.2: The Radon transform of f at (t, θ) is the line integral of f along line L.

Chapter 18

Transmission Tomography II

18.1 Chapter Summary

According to the Central Slice Theorem, if we have all the line integrals through the attenuation function $f(x, y)$ then we have the two-dimensional Fourier transform of $f(x, y)$. To get $f(x, y)$ we need to invert the two-dimensional Fourier transform; that is the topic of this chapter.

18.2 Inverting the Fourier Transform

The Fourier-transform inversion formula for two-dimensional functions tells us that the function $f(x, y)$ can be obtained as

$$f(x, y) = \frac{1}{4\pi^2} \int \int F(u, v) e^{-i(xu+yv)} du dv. \tag{18.1}$$

We now derive alternative inversion formulas.

18.2.1 Back-Projection

Let $g(\theta, t)$ be any function of the variables θ and t; for example, it could be the Radon transform. As with the Radon transform, we imagine that each pair (θ, t) corresponds to one line through the x, y-plane. For each fixed point (x, y) we assign to this point the sum of the quantities $g(\theta, t)$ for every pair (θ, t) such that the point (x, y) lies on the associated line. The summing process is integration and the *back-projection* function at (x, y) is

$$BP_g(x, y) = \int g(\theta, x\cos\theta + y\sin\theta)d\theta.$$

The operation of back-projection will play an important role in what follows in this chapter.

18.2.2 Ramp Filter, then Back-project

Expressing the double integral in Equation (18.1) in polar coordinates (ω, θ), with $\omega \geq 0$, $u = \omega\cos\theta$, and $v = \omega\sin\theta$, we get

$$f(x, y) = \frac{1}{4\pi^2}\int_0^{2\pi}\int_0^{\infty} F(u, v)e^{-i(xu+yv)}\omega d\omega d\theta,$$

or

$$f(x, y) = \frac{1}{4\pi^2}\int_0^{\pi}\int_{-\infty}^{\infty} F(u, v)e^{-i(xu+yv)}|\omega| d\omega d\theta.$$

Now write

$$F(u, v) = F(\omega\cos\theta, \omega\sin\theta) = R_f(\theta, \omega),$$

where $R_f(\theta, \omega)$ is the FT with respect to t of $r_f(\theta, t)$, so that

$$\int_{-\infty}^{\infty} F(u, v)e^{-i(xu+yv)}|\omega| d\omega = \int_{-\infty}^{\infty} R_f(\theta, \omega)|\omega| e^{-i\omega t} d\omega.$$

The function $g_f(\theta, t)$ defined for $t = x\cos\theta + y\sin\theta$ by

$$g_f(\theta, x\cos\theta + y\sin\theta) = \frac{1}{2\pi}\int_{-\infty}^{\infty} R_f(\theta, \omega)|\omega| e^{-i\omega t} d\omega \tag{18.2}$$

is the result of a linear filtering of $r_f(\theta, t)$ using a *ramp filter* with transfer function $H(\omega) = |\omega|$. Then,

$$f(x, y) = \frac{1}{2\pi}\int_0^{\pi} g_f(\theta, x\cos\theta + y\sin\theta)d\theta \tag{18.3}$$

gives $f(x, y)$ as the result of a *back-projection operator*; for every fixed value of (θ, t) add $g_f(\theta, t)$ to the current value at the point (x, y) for all (x, y)

lying on the straight line determined by θ and t by $t = x\cos\theta + y\sin\theta$. The final value at a fixed point (x, y) is then the average of all the values $g_f(\theta, t)$ for those (θ, t) for which (x, y) is on the line $t = x\cos\theta + y\sin\theta$. It is therefore said that $f(x, y)$ can be obtained by *filtered back-projection* (FBP) of the line-integral data.

Knowing that $f(x, y)$ is related to the complete set of line integrals by filtered back-projection suggests that, when only finitely many line integrals are available, a similar ramp filtering and back-projection can be used to estimate $f(x, y)$; in the clinic this is the most widely used method for the reconstruction of tomographic images.

18.2.3 Back-project, then Ramp Filter

There is a second way to recover $f(x, y)$ using back-projection and filtering, this time in the reverse order; that is, we back-project the Radon transform and then ramp filter the resulting function of two variables. We begin again with the relation

$$f(x,y) = \frac{1}{4\pi^2}\int_0^{2\pi}\int_0^{\infty} F(u,v)e^{-i(xu+yv)}\omega d\omega d\theta,$$

which we write as

$$f(x,y) = \frac{1}{4\pi^2}\int_0^{2\pi}\int_0^{\infty} \frac{F(u,v)}{\sqrt{u^2+v^2}}\sqrt{u^2+v^2}e^{-i(xu+yv)}\omega d\omega d\theta$$

$$= \frac{1}{4\pi^2}\int_0^{2\pi}\int_0^{\infty} G(u,v)\sqrt{u^2+v^2}e^{-i(xu+yv)}\omega d\omega d\theta, \tag{18.4}$$

using

$$G(u,v) = \frac{F(u,v)}{\sqrt{u^2+v^2}}$$

for $(u, v) \neq (0, 0)$. Equation (18.4) expresses $f(x, y)$ as the result of performing a two-dimensional ramp filtering of $g(x, y)$, the inverse Fourier transform of $G(u, v)$. We show now that $g(x, y)$ is the back-projection of the function $r_f(\theta, t)$; that is, we show that

$$g(x,y) = \frac{1}{2\pi}\int_0^{\pi} r_f(\theta, x\cos\theta + y\sin\theta)d\theta.$$

We have

$$g(x,y) = \frac{1}{4\pi^2}\int_0^{\pi}\int_{-\infty}^{\infty} G(\omega\cos\theta, \omega\sin\theta)|\omega|e^{-i\omega(x\cos\theta+y\sin\theta)}d\omega d\theta$$

$$= \frac{1}{4\pi^2}\int_0^{\pi}\int_{-\infty}^{\infty} F(\omega\cos\theta, \omega\sin\theta)e^{-i\omega(x\cos\theta + y\sin\theta)}d\omega d\theta$$

$$= \frac{1}{4\pi^2}\int_0^{\pi}\int_{-\infty}^{\infty} R_f(\theta,\omega)e^{-i\omega(x\cos\theta + y\sin\theta)}d\omega d\theta$$

$$= \frac{1}{2\pi}\int_0^{\pi} r_f(\theta, x\cos\theta + y\sin\theta)d\theta,$$

as required.

18.2.4 Radon's Inversion Formula

To get Radon's inversion formula, we need two basic properties of the Fourier transform. First, if $f(x)$ has Fourier transform $F(\gamma)$ then the derivative $f'(x)$ has Fourier transform $-i\gamma F(\gamma)$. Second, if $F(\gamma) = \text{sgn}(\gamma)$, the function that is $\frac{\gamma}{|\gamma|}$ for $\gamma \neq 0$, and equal to zero for $\gamma = 0$, then its inverse Fourier transform is $f(x) = \frac{1}{i\pi x}$.

Writing equation (18.2) as

$$g_f(\theta, t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \omega R_f(\theta,\omega)\text{sgn}(\omega)e^{-i\omega t}d\omega,$$

we see that g_f is the inverse Fourier transform of the product of the two functions $\omega R_f(\theta,\omega)$ and $\text{sgn}(\omega)$. Consequently, g_f is the convolution of their individual inverse Fourier transforms, $i\frac{\partial}{\partial t}r_f(\theta,t)$ and $\frac{1}{i\pi t}$; that is,

$$g_f(\theta,t) = \frac{1}{\pi}\int_{-\infty}^{\infty}\frac{\partial}{\partial t}r_f(\theta,s)\frac{1}{t-s}ds,$$

which is the Hilbert transform of the function $\frac{\partial}{\partial t}r_f(\theta,t)$, with respect to the variable t. Radon's inversion formula is then

$$f(x,y) = \frac{1}{2\pi}\int_0^{\pi} HT(\frac{\partial}{\partial t}r_f(\theta,t))d\theta.$$

18.3 From Theory to Practice

What we have just described is the theory. What happens in practice?

18.3.1 The Practical Problems

Of course, in reality we never have the Radon transform $r_f(\theta, t)$ for all values of its variables. Only finitely many angles θ are used, and, for each θ, we will have (approximate) values of line integrals for only finitely many t. Therefore, taking the Fourier transform of $r_f(\theta, t)$, as a function of the single variable t, is not something we can actually do. At best, we can approximate $R_f(\theta, \omega)$ for finitely many θ. From the Central Slice Theorem, we can then say that we have approximate values of $F(\omega \cos\theta, \omega \sin\theta)$, for finitely many θ. This means that we have (approximate) Fourier transform values for $f(x, y)$ along finitely many lines through the origin, like the spokes of a wheel. The farther from the origin we get, the fewer values we have, so the *coverage* in Fourier space is quite uneven. The low-spatial-frequencies are much better estimated than higher ones, meaning that we have a low-pass version of the desired $f(x, y)$. The filtered back-projection approaches we have just discussed both involve ramp filtering, in which the higher frequencies are increased, relative to the lower ones. This too can only be implemented approximately, since the data is noisy and careless ramp filtering will cause the reconstructed image to be unacceptably noisy.

18.3.2 A Practical Solution: Filtered Back-Projection

We assume, to begin with, that we have finitely many line integrals, that is, we have values $r_f(\theta, t)$ for finitely many θ and finitely many t. For each fixed θ we estimate the Fourier transform, $R_f(\theta, \omega)$. This step can be performed in various ways, and we can freely choose the values of ω at which we perform the estimation. The FFT will almost certainly be involved in calculating the estimates of $R_f(\theta, \omega)$.

For each fixed θ we multiply our estimated values of $R_f(\theta, \omega)$ by $|\omega|$ and then use the FFT again to inverse Fourier transform, to achieve a ramp filtering of $r_f(\theta, t)$ as a function of t. Note, however, that when $|\omega|$ is large, we may multiply by a smaller quantity, to avoid enhancing noise. We do this for each angle θ, to get a function of (θ, t), which we then back-project to get our final image. This is ramp-filtering, followed by back-projection, as applied to the finite data we have.

It is also possible to mimic the second approach to inversion, that is, to back-project onto the pixels each $r_f(\theta, t)$ that we have, and then to perform a ramp filtering of this two-dimensional array of numbers to obtain the final image. In this case, the two-dimensional ramp filtering involves many applications of the FFT.

There is a third approach. Invoking the Central Slice Theorem, we can say that we have finitely many approximate values of $F(u, v)$, the Fourier transform of the attenuation function $f(x, y)$, along finitely many lines through the origin. The first step is to use these values to estimate the

values of $F(u, v)$ at the points of a rectangular grid. This step involves *interpolation* [249, 254]. Once we have (approximate) values of $F(u, v)$ on a rectangular grid, we perform a two-dimensional FFT to obtain our final estimate of the (discretized) $f(x, y)$.

18.4 Some Practical Concerns

As computer power increases and scanners become more sophisticated, there is pressure to include more dimensionality in the scans. This means going beyond slice-by-slice tomography to fully three-dimensional images, or even including time as the fourth dimension, to image dynamically. This increase in dimensionality comes at a cost, however [236]. Besides the increase in radiation to the patient, there are other drawbacks, such as longer acquisition time, storing large amounts of data, processing and analyzing this data, displaying the results, reading and understanding the higher-dimensional images, and so on.

18.5 Summary

We have seen how the problem of reconstructing a function from line integrals arises in transmission tomography. The Central Slice Theorem connects the line integrals and the Radon transform to the Fourier transform of the desired attenuation function. Various approaches to implementing the Fourier Inversion Formula lead to filtered back-projection algorithms for the reconstruction. In x-ray tomography, as well as in PET, viewing the data as line integrals ignores the statistical aspects of the problem, and in SPECT, it ignores, as well, the important physical effects of attenuation. To incorporate more of the physics of the problem, iterative algorithms based on statistical models have been developed. We shall consider some of these algorithms later.

Chapter 19

Emission Tomography

19.1 Chapter Summary

In this chapter we describe the two modalities of emission tomography, *positron emission tomography* (PET) and *single photon emission computed tomography* (SPECT), and introduce the basic mathematical models for both.

The man in Figure 19.1 is the mathematician Paul Dirac, often called "the British Einstein" . Almost all cancer survivors have had a PET scan, a marvelous invention that owes its existence to the genius of this man. Those who knew him often remarked on his "strangeness" ; recent studies have suggested that both Dirac and his father were autistic.

This is Dirac's Equation from quantum mechanics, which predicted the existence of the positron and eventually led to PET scans:

$$i\hbar\frac{\partial\psi}{\partial t} = \frac{\hbar c}{i}\Big(\alpha_1\frac{\partial\psi}{\partial x_1} + \alpha_2\frac{\partial\psi}{\partial x_2} + \alpha_3\frac{\partial\psi}{\partial x_3}\Big) + \alpha_4 mc^2\psi.$$

In 1930 Dirac added his equation, now inscribed on the wall of Westminster Abbey, to the developing field of quantum mechanics. This equation agreed remarkably well with experimental data on the behavior of electrons in electric and magnetic fields, but it also seemed to allow for nonsensical solutions, such as spinning electrons with negative energy.

The next year, Dirac realized that what the equation was calling for was *anti-matter*, a particle with the same mass as the electron, but with

a positive charge. In the summer of 1932, Carl Anderson, working at Cal Tech, presented clear evidence for the existence of such a particle, which we now call the *positron*. What seemed like the height of science fiction in 1930 has become commonplace today.

Both PET and SPECT scans rely on metabolism and so must be performed on living beings, principally people and small animals. The pig in Figure 19.2 is having his heart imaged using SPECT, as part of a research effort to study the effectiveness of certain SPECT imaging algorithms. The hearts of pigs are similar to our own, which makes the pig a good subject for this study.

19.2 Positron Emission Tomography

As we noted previously, detection in the PET case means the recording of two photons at nearly the same time at two different detectors. The locations of these two detectors then provide the end points of the line segment passing, more or less, through the site of the original positron emission. Therefore, each possible pair of detectors determines a *line of response* (LOR). When a LOR is recorded, it is assumed that a positron was emitted somewhere along that line. The PET data consists of a chronological list of LOR that are recorded.

Let the LOR be parameterized by the variable s, with $s = 0$ and $s = c$ denoting the two ends, and c the distance from one end to the other. For a fixed value $s = s_0$, let $P(s)$ be the probability of reaching s for a photon resulting from an emission at s_0. For small $\Delta s > 0$ the probability that a photon that reached s is absorbed in the interval $[s, s+\Delta s]$ is approximately $\mu(s)\Delta s$, where $\mu(s) \geq 0$ is the photon attenuation density at s. Then $P(s + \Delta s) \approx P(s)[1 - \mu(s)\Delta s]$, so that

$$P(s + \Delta s) - P(s) \approx -P(s)\mu(s)\Delta s.$$

Dividing by Δs and letting Δs go to zero, we get

$$P'(s) = -P(s)\mu(s).$$

It follows that

$$P(s) = e^{-\int_{s_0}^{s} \mu(t)dt}.$$

The probability that the photon will reach $s = c$ and be detected is then

$$P(c) = e^{-\int_{s_0}^{c} \mu(t)dt}.$$

Similarly, we find that the probability that a photon will succeed in reaching $s = 0$ from s_0 is

$$P(0) = e^{-\int_0^{s_0} \mu(t)dt}.$$

Since having one photon reach $s = 0$ and the other reach $s = c$ are independent events, their probabilities multiply, so that the probability that both photons reach their destinations and a coincident detection is recorded for this LOR is

$$e^{-\int_0^c \mu(t)dt}.$$

The expected number of coincident detections along the LOR is then proportional to

$$\int_0^c f(s)e^{-\int_0^c \mu(t)dt}ds = e^{-\int_0^c \mu(t)dt} \int_0^c f(s)ds, \tag{19.1}$$

where $f(s)$ is the intensity of radionuclide at s.

Let y_i be the number of coincidence detections associated with the ith LOR. If we are willing to equate the actual count with the expected count, and assuming we know the attenuation function $\mu(s)$, we can estimate the line integral $\int_0^c f(s)ds$ along the ith LOR as

$$\int_0^c f(s)ds = y_i e^{\int_0^c \mu(t)dt}.$$

So, once again, we have line-integral data pertaining to the function of interest.

19.3 Single-Photon Emission Tomography

We turn now to single-photon computed emission tomography (SPECT).

19.3.1 Sources of Degradation to be Corrected

We remarked earlier that there are at least three degradations that need to be corrected before the line-integral model and FBP can be successfully applied in the SPECT case [181]: attenuation, scatter, and spatially dependent resolution. There are mathematical ways to correct for both spatially varying resolution and uniform attenuation [243]. Correcting for the more realistic non-uniform and patient-specific attenuation is more difficult and is the subject of on-going research.

Some photons never reach the detectors because they are absorbed in

the body. As in the PET case, correcting for attenuation requires knowledge of the patient's body; this knowledge can be obtained by performing a transmission scan at the same time. In contrast to the PET case, the attenuation due to absorption is more difficult to correct, since it does not involve merely the line integral of the attenuation function, but a half-line integral that depends on the distribution of matter between each photon source and each detector.

While some photons are absorbed within the body, others are first deflected and then detected; this is called *scatter*. Consequently, some of the detected photons do not come from where they seem to come from. The scattered photons often have reduced energy, compared to *primary*, or non-scattered, photons, and scatter correction can be based on this energy difference; see [181].

Finally, even if there were no attenuation and no scatter, it would be incorrect to view the detected photons as having originated along a single straight line from the detector. Due to the use of a lead collimator in SPECT (but not in PET), the detectors have a cone of acceptance that widens as it recedes from the detector. This results in spatially varying resolution.

It is not uncommon, however, to make the simplifying assumption that all photons detected at a given detector originated along a single line. As in the PET case previously discussed, the probability that a photon emitted at the point on the line corresponding to the variable $s = s_0$ will reach $s = c$ and be detected is then

$$P(s_0) = e^{-\int_{s_0}^{c} \mu(t)dt}.$$

If $f(s)$ is the expected number of photons emitted from point s during the scanning, then the expected number of photons detected at c and originating along this line is proportional to

$$\int_0^c f(s)e^{-\int_s^c \mu(t)dt}ds. \tag{19.2}$$

Notice the difference between the integral in Equation (19.2) and the one in Equation (19.1).

The integral in Equation (19.2) varies with the line being considered; the resulting function of lines is called the *attenuated Radon transform*.

If the attenuation function μ is constant, then the attenuated Radon transform is called the *exponential Radon transform*. Since

$$\int_s^c \mu dt = \mu(c - s),$$

the integral in (19.2) is now

$$e^{-\mu c}\int_0^c f(s)e^{\mu s}ds = e^{-\mu c}\int_0^\infty f(s)e^{-(-\mu)s}ds = e^{-\mu c}\mathcal{F}(-\mu),$$

where $\mathcal{F}$ denotes the Laplace transform of f. Since the function $f(s)$ is zero outside a bounded interval, we may safely assume that the Laplace transform is defined for all real values of the argument.

In practice, one sometimes assumes, initially, that $\mu = 0$ and that the counts at each detector are essentially integrals of f along a single line. Filtered back-projection is then used to reconstruct an image. Since the image does not reflect the effects of attenuation, it can be "corrected" during the back-projection phase.

Spatially varying resolution complicates the quantitation problem, which is the effort to determine the exact amount of radionuclide present within a given region of the body, by introducing the *partial volume effect* and *spill-over* (see [263]). To a large extent, these problems are shortcomings of reconstruction based on the line-integral model. If we assume that all photons detected at a particular detector came from points within a narrow strip perpendicular to the camera face, and we reconstruct the image using this assumption, then photons coming from locations outside this strip will be incorrectly attributed to locations within the strip (spill-over), and therefore not correctly attributed to their true source location. If the true source location also has its counts raised by spill-over, the net effect may not be significant; if, however, the true source is a hot spot surrounded by cold background, it gets no spill-over from its neighbors and its true intensity value is underestimated, resulting in the partial-volume effect. The term "partial volume" indicates that the hot spot is smaller than the region that the line-integral model offers as the source of the emitted photons. One way to counter these effects is to introduce a description of the spatially dependent blur into the reconstruction, which is then performed by iterative methods [220].

In the SPECT case, as in most such inverse problems, there is a trade-off to be made between careful modeling of the physical situation and computational tractability. The FBP method slights the physics in favor of computational simplicity and speed. In recent years, iterative methods that incorporate more of the physics have become competitive.

19.3.2 The Discrete Model

In iterative reconstruction we begin by *discretizing* the problem; that is, we imagine the region of interest within the patient to consist of finitely many tiny squares, called *pixels* for two-dimensional processing or cubes, called *voxels* for three-dimensional processing. In what follows we shall not distinguish the two cases, but as a linguistic shorthand, we shall refer to 'pixels' indexed by $j = 1, ..., J$. The detectors are indexed by $i = 1, ..., I$, the count obtained at detector i is denoted y_i, and the vector $\mathbf{y} = (y_1, ..., y_I)^T$ is our data. In practice, for the fully three-dimensional case, I and J can be several hundred thousand.

We imagine that each pixel j has its own level of concentration of radioactivity and these concentration levels are what we want to determine. Proportional to these concentration levels are the average rates of emission of photons; the average rate for j we denote by x_j. The goal is to determine the vector $\mathbf{x} = (x_1, ..., x_J)^T$ from $\mathbf{y}$.

19.3.3 Discrete Attenuated Radon Transform

To achieve our goal we must construct a model that relates $\mathbf{y}$ to $\mathbf{x}$. One way to do that is to discretize the attenuated Radon Transform [156, 250].

The objective is to describe the contribution to the count data from the intensity x_j at the jth pixel. We assume, for the moment, that all the radionuclide is concentrated within the jth pixel, and we compute the resulting attenuated Radon Transform. Following [156, 250], we adopt a ray model for detection, which means that corresponding to each detector is a line of acceptance and that all the counts recorded at that detector came from pixels that intersect this line. This is a simplification, of course, since each detector has a solid angle of acceptance, which leads to depth-dependent blur.

For notational simplicity, we suppose that the line of acceptance associated with the ith detector is parameterized by arc-length $s \geq 0$, with $s = c > 0$ corresponding to the point closest to the detector, within the body, $s = 0$ corresponding to the point farthest from the detector, at which the line leaves the body, $s = b < c$ the point closest to the detector within the jth pixel, and $s = a < b$ the point farthest from the detector at which the line leaves the jth pixel. The length of the intersection of the jth pixel with the line is then $d_{ij} = b - a$.

We are assuming that all the radionuclide is within the jth pixel, with intensity distribution (proportional to) x_j, so the value at detector i of the attenuated Radon Transform is

$$A_{ij} = \int_a^b x_j e^{-\int_s^c \mu(t)dt} ds. \tag{19.3}$$

We assume that the attenuation is uniformly equal to $\mu_j \geq 0$ within the jth pixel, so we can write

$$A_{ij} = \int_a^b x_j e^{-\int_s^b \mu_j dt - \int_b^c \mu(t)dt} ds,$$

or

$$A_{ij} = x_j e^{-\int_b^c \mu(t)dt} \int_a^b e^{(s-b)\mu_j} ds.$$

If $\mu_j = 0$, then we have

$$A_{ij} = x_j e^{-\int_b^c \mu(t)dt} d_{ij},$$

while if $\mu_j > 0$ we have

$$A_{ij} = \left(x_j e^{-\int_b^c \mu(t)dt} d_{ij} \right) S_{ij},$$

where

$$S_{ij} = \frac{1}{d_{ij}} \int_a^b e^{(b-s)\mu_j} ds = \frac{1}{\mu_j d_{ij}} (1 - e^{-\mu_j d_{ij}}).$$

We can then write

$$A_{ij} = x_j W_{ij},$$

for each j and i.

Since the function

$$g(t) = \frac{1}{t}(1 - e^{-t})$$

is positive for positive t, $g(0) = 1$, and $g(+\infty) = 0$, it is reasonable to view S_{ij} as the survival proportion associated with the jth pixel and the line from the ith detector. Expanding the exponential in S_{ij} in a power series, we find that

$$S_{ij} = \frac{1}{\mu_j d_{ij}} (1 - e^{-\mu_j d_{ij}}) \approx 1 - \frac{1}{2} \mu_j d_{ij},$$

so that the loss proportion is approximately $\frac{1}{2}\mu_j d_{ij}$. If we were to adopt the decaying exponential model for a photon surviving its passage through the jth pixel, and assume all the radionuclide was initially at the far side of the jth pixel, we would replace S_{ij} with $e^{-\mu_j d_{ij}}$, which is approximately $1 - \mu_j d_{ij}$, so that the loss proportion is approximately $\mu_j d_{ij}$. This is twice the loss proportion that we got using the other model, and is larger because we are assuming that all the photons emitted from within the jth pixel have to attempt to travel through the entire jth pixel, whereas, due to the spreading of the radionuclide throughout the pixel, the average photon journey through the pixel is only half of the length d_{ij}.

Having found the values W_{ij}, we form the matrix W having these entries and then find a non-negative solution of the system of equations $Wx = y$, using one of a number of iterative algorithms, including the EMML. Contrary to what is stated in [250], it may not be appropriate to consider W_{ij} as the probability that a photon emitted at the jth pixel is detected at the ith detector, even though $0 \leq W_{ij} \leq 1$ for each i and j. If viewed that way, it would be the case that

$$\sum_{i=1}^{I} W_{ij}$$

would be the probability of detecting a photon emitted from the jth pixel; we have no guarantee, however, that this sum is not greater than one.

It is significant that the authors in [250] realize that the EMML iterative

algorithm can be used to find a non-negative solution of $Wx = y$, even though no stochastic model for the data is assumed in their derivation. Their development involves discretizing the attenuated Radon Transform, which involves no randomness, and viewing the count data as approximate values of this discrete function.

There is another approach that can be used to relate the count data to the intensity levels x_j. This other approach is based on a stochastic model, as we describe next.

19.3.4 A Stochastic Model

Another way to relate the count data to the intensities x_j is to adopt the model of *independent Poisson emitters*. For $i = 1, ..., I$ and $j = 1, ..., J$, denote by Z_{ij} the random variable whose value is to be the number of photons emitted from pixel j, and detected at detector i, during the scanning time. We assume that the members of the collection $\{Z_{ij} | i = 1, ..., I, j = 1, ..., J\}$ are independent. In keeping with standard practice in modeling radioactivity, we also assume that the Z_{ij} are Poisson-distributed.

Generally, the signal-to-noise ratio (SNR) is the ratio of the mean of a distribution to its standard deviation (the square root of the variance). In the case of the Poisson distribution, the variance and the mean are the same, so the SNR is the square root of the mean; therefore, the higher the mean the higher the SNR.

We assume that Z_{ij} is a Poisson random variable whose mean value (and variance) is $\lambda_{ij} = P_{ij}x_j$. Here the $x_j \geq 0$ is the average rate of emission from pixel j, as discussed previously, and $P_{ij} \geq 0$ is the probability that a photon emitted from pixel j will be detected at detector i. The calculation of the P_{ij} can be quite similar to the derivation of the W_{ij} in the previous subsection, with the exception that we do need to have

$$\sum_{i=1}^{I} P_{ij} \leq 1.$$

We then define the random variables $Y_i = \sum_{j=1}^{J} Z_{ij}$, the total counts to be recorded at detector i; our actual count y_i is then the observed value of the random variable Y_i. Note that the actual values of the individual Z_{ij} are not observable.

Any Poisson-distributed random variable has a mean equal to its variance. The *signal-to-noise ratio* (SNR) is usually taken to be the ratio of the mean to the standard deviation, which, in the Poisson case, is then the square root of the mean. Consequently, the Poisson SNR increases as the mean value increases, which points to the desirability (at least, statistically speaking) of higher dosages to the patient.

Having found the P_{ij}, we take P to be the matrix with these entries.

Since Px is the vector of expected counts at the various detectors, and y is the vector of actual counts, trying to find a non-negative solution of the system $y = Px$ may not seem completely reasonable. However, this is what several well known iterative algorithms do, even ones such as the EMML that were not originally designed for this purpose.

19.3.5 Reconstruction as Parameter Estimation

The goal is to estimate the distribution of radionuclide intensity by calculating the vector $\mathbf{x}$. The entries of $\mathbf{x}$ are parameters and the data are instances of random variables, so the problem looks like a fairly standard parameter estimation problem of the sort studied in beginning statistics. One of the basic tools for statistical parameter estimation is likelihood maximization, which is playing an increasingly important role in medical imaging. There are several problems, however. One is that the number of parameters is quite large, as large as the number of data values, in most cases. Standard statistical parameter estimation usually deals with the estimation of a handful of parameters. Another problem is that we do not know what the P_{ij} are. These values will vary from one patient to the next, since whether or not a photon makes it from a given pixel to a given detector depends on the geometric relationship between detector i and pixel j, as well as what is in the patient's body between these two locations. If there are ribs or skull getting in the way, the probability of making it goes down. If there are just lungs, the probability goes up. These values can change during the scanning process, when the patient moves. Some motion is unavoidable, such as breathing and the beating of the heart. Determining good values of the P_{ij} in the absence of motion, and correcting for the effects of motion, are important parts of SPECT image reconstruction.

19.4 Relative Advantages

In [212], Ollinger and Fessler discuss some of the relative advantages of these two modes of emission tomography.

Attenuation, which is primarily the scattering of photons by the body to locations outside the field of view of the detecting cameras, is harder to correct in SPECT. The radiopharmaceuticals used in SPECT must incorporate heavy isotopes, such as thallium and technetium; since these do not occur naturally in biologically active molecules, the synthesis of physiologically useful tracers is a challenge. In contrast, in PET the positron-emitting

isotopes of carbon, nitrogen, oxygen and fluorine that are used occur naturally in many compounds of biological interest and can therefore be easily incorporated into useful radiopharmaceuticals.

Because collimation is performed by the computer in PET, while SPECT must employ lead collimators, which absorb many of the photons, the sensitivity of the detecting gamma cameras in SPECT is reduced, in comparison to PET.

On the other side of the balance sheet, the short half-life of most positron-emitting isotopes necessitates an on-site cyclotron, while the isotopes used in SPECT have longer half-lives and can be stored. Also, the scanners for PET are more expensive than those used in SPECT.

At any given time, computer speed limits the size of the problem that can be dealt with. While 2D reconstructions are clinically feasible, fully 3D imaging (not to mention dynamic, 4D imaging) poses more of a challenge, hence the need for continuing algorithm development.

FIGURE 19.1: Paul Dirac: his equation predicted positrons.

FIGURE 19.2: A pet getting a PET scan? Not quite.

Chapter 20

Magnetic Resonance Imaging

20.1 Chapter Summary

In elements with an odd number of protons, such as hydrogen, the nucleus itself will have a net magnetic moment. The objective in *magnetic resonance imaging* (MRI) is to determine the density of such elements in a volume of interest within the body. This is achieved by forcing the individual spinning nuclei to emit signals that, while too weak to be detected alone, are detectable in the aggregate. Fourier-transform estimation and extrapolation techniques play a major role in the rapidly expanding field of magnetic resonance imaging [265, 157].

The gentleman in Figure 20.1 is Raymond Damadian, who received the Nobel prize in 2004 for inventing MRI. In Figure 20.2 we see a typical MRI image of a head; check out the eyeballs.

20.2 Slice Isolation

When the external magnetic field is the *static field* $B_0\mathbf{k}$, that is, the magnetic field has strength B_0 and axis $\mathbf{k} = (0, 0, 1)$, then the Larmor frequency is the same everywhere and equals $\omega_0 = \gamma B_0$, where γ is the gyromagnetic constant. If, instead, we impose an external magnetic field $(B_0 + G_z(z - z_0))\mathbf{k}$, for some constant G_z, then the Larmor frequency is ω_0 only within the plane $z = z_0$. This external field now includes a *gradient field*.

20.3 Tipping

When a magnetic dipole moment that is aligned with $\mathbf{k}$ is given a component in the x, y-plane, it begins to precess around the z-axis, with frequency equal to its Larmor frequency. To create this x, y-plane component, we apply a *radio-frequency field* (rf field)

$$H_1(t)(\cos(\omega t)\mathbf{i} + \sin(\omega t)\mathbf{j}).$$

The function $H_1(t)$ typically lasts only for a short while, and the effect of imposing this rf field is to tip the aligned magnetic dipole moment axes away from the z-axis, initiating precession. Those dipole axes that tip most are those whose Larmor frequency is ω. Therefore, if we first isolate the slice $z = z_0$ and then choose $\omega = \omega_0$, we tip primarily those dipole axes within the plane $z = z_0$. The dipoles that have been tipped ninety degrees into the x, y-plane generate the strongest signal. How much tipping occurs also depends on $H_1(t)$, so it is common to select $H_1(t)$ to be constant over the time interval $[0, \tau]$, and zero elsewhere, with integral $\frac{\pi}{2\gamma}$. This $H_1(t)$ is called a $\frac{\pi}{2}$-pulse, and tips those axes with Larmor frequency ω_0 into the x, y-plane.

20.4 Imaging

The information we seek about the proton density function is contained within the received signal. By carefully adding gradient fields to the external field, we can make the Larmor frequency spatially varying, so that each

frequency component of the received signal contains a piece of the information we seek. The proton density function is then obtained through Fourier transformations.

20.4.1 The Line-Integral Approach

Suppose that we have isolated the plane $z = z_0$ and tipped the aligned axes using a $\frac{\pi}{2}$-pulse. After the tipping has been completed, we introduce an external field $(B_0 + G_x x)\mathbf{k}$, so that now the Larmor frequency of dipoles within the plane $z = z_0$ is $\omega(x) = \omega_0 + \gamma G_x x$, which depends on the x-coordinate of the point. The result is that the component of the received signal associated with the frequency $\omega(x)$ is due solely to those dipoles having that x coordinate. Performing an FFT of the received signal gives us line integrals of the density function along lines in the x, y-plane having fixed x-coordinate.

More generally, if we introduce an external field $(B_0 + G_x x + G_y y)\mathbf{k}$, the Larmor frequency is constant at $\omega(x, y) = \omega_0 + \gamma(G_x x + G_y y) = \omega_0 + \gamma s$ along lines in the x, y-plane with equation

$$G_x x + G_y y = s.$$

Again performing an FFT on the received signal, we obtain the integral of the density function along these lines. In this way, we obtain the three-dimensional Radon transform of the desired density function. The central slice theorem for this case tells us that we can obtain the Fourier transform of the density function by performing a one-dimensional Fourier transform with respect to the variable s. For each fixed (G_x, G_y) we obtain this Fourier transform along a ray through the origin. By varying the (G_x, G_y) we get the entire Fourier transform. The desired density function is then obtained by Fourier inversion.

20.4.2 Phase Encoding

In the line-integral approach, the line-integral data is used to obtain values of the Fourier transform of the density function along lines through the origin in Fourier space. It would be more convenient to have Fourier-transform values on the points of a rectangular grid. We can obtain this by selecting the gradient fields to achieve *phase encoding*.

Suppose that, after the tipping has been performed, we impose the external field $(B_0 + G_y y)\mathbf{k}$ for T seconds. The effect is to alter the precession frequency from ω_0 to $\omega(y) = \omega_0 + \gamma G_y y$. A harmonic $e^{i\omega_0 t}$ is changed to

$$e^{i\omega_0 t} e^{i\gamma G_y y t},$$

so that, after T seconds,we have

$$e^{i\omega_0 T} e^{i\gamma G_y y T}.$$

For $t \geq T$, the harmonic $e^{i\omega_0 t}$ returns, but now it is

$$e^{i\omega_0 t} e^{i\gamma G_y y T}.$$

The effect is to introduce a phase shift of $\gamma G_y y T$. Each point with the same y-coordinate has the same phase shift.

After time T, when this gradient field is turned off, we impose a second external field, $(B_0 + G_x x)\mathbf{k}$. Because this gradient field alters the Larmor frequencies, at times $t \geq T$ the harmonic $e^{i\omega_0 t} e^{i\gamma G_y y T}$ is transformed into

$$e^{i\omega_0 t} e^{i\gamma G_y y T} e^{i\gamma G_x x t}.$$

The received signal is now

$$S(t) = e^{i\omega_0 t} \int \int \rho(x, y) e^{i\gamma G_y y T} e^{i\gamma G_x x t} dx dy,$$

where $\rho(x, y)$ is the value of the proton density function at (x, y). Removing the $e^{i\omega_0 t}$ factor, we have

$$\int \int \rho(x, y) e^{i\gamma G_y y T} e^{i\gamma G_x x t} dx dy,$$

which is the Fourier transform of $\rho(x, y)$ at the point $(\gamma G_x t, \gamma G_y T)$. By selecting equi-spaced values of t and altering the G_y, we can get the Fourier transform values on a rectangular grid.

20.5 The General Formulation

The external magnetic field generated in the MRI scanner is generally described by

$$H(r, t) = (H_0 + \mathbf{G}(t) \cdot \mathbf{r})\mathbf{k} + H_1(t)(\cos(\omega t)\mathbf{i} + \sin(\omega t)\mathbf{j}). \tag{20.1}$$

The vectors $\mathbf{i}, \mathbf{j}$, and $\mathbf{k}$ are the unit vectors along the coordinate axes, and $\mathbf{r} = (x, y, z)$. The vector-valued function $\mathbf{G}(t) = (G_x(t), G_y(t), G_z(t))$ produces the *gradient field*

$$\mathbf{G}(t) \cdot \mathbf{r}.$$

The magnetic field component in the x, y plane is the *radio frequency* (rf) field.

If $\mathbf{G}(t) = 0$, then the Larmor frequency is ω_0 everywhere. Using $\omega = \omega_0$ in the rf field, with a $\frac{\pi}{2}$-pulse, will then tip the aligned axes into the x, y-plane and initiate precession. If $\mathbf{G}(t) = \theta$, for some direction vector θ, then the Larmor frequency is constant on planes $\theta \cdot \mathbf{r} = s$. Using an rf field with frequency $\omega = \gamma(H_0 + s)$ and a $\frac{\pi}{2}$-pulse will then tip the axes in this plane into the x, y-plane. The strength of the received signal will then be proportional to the integral, over this plane, of the proton density function. Therefore, the measured data will be values of the three-dimensional Radon transform of the proton density function, which is related to its three-dimensional Fourier transform by the Central Slice Theorem. Later, we shall consider two more widely used examples of $\mathbf{G}(t)$.

20.6 The Received Signal

We assume now that the function $H_1(t)$ is a *short* $\frac{\pi}{2}$*-pulse*, that is, it has constant value over a short time interval $[0, \tau]$ and has integral $\frac{\pi}{2\gamma}$. The received signal produced by the precessing magnetic dipole moments is approximately

$$S(t) = \int_{R^3} \rho(\mathbf{r}) \exp(-i\gamma(\int_0^t \mathbf{G}(s)ds) \cdot \mathbf{r}) \exp(-t/T_2)d\mathbf{r}, \tag{20.2}$$

where $\rho(\mathbf{r})$ is the proton density function, and T_2 is the *transverse* or *spin-spin* relaxation time. The vector integral in the exponent is

$$\int_0^t \mathbf{G}(s)ds = (\int_0^t G_x(s)ds, \int_0^t G_y(s)ds, \int_0^t G_z(s)ds).$$

Now imagine approximating the function $G_x(s)$ over the interval $[0, t]$ by a step function that is constant over small subintervals, that is, $G_x(s)$ is approximately $G_x(n\Delta)$ for s in the interval $[n\Delta, (n+1)\Delta)$, with $n = 1, ..., N$ and $\Delta = \frac{t}{N}$. During the interval $[n\Delta, (n+1)\Delta)$, the presence of this gradient field component causes the phase to change by the amount $x\gamma G_x(n\Delta)\Delta$, so that by the time we reach $s = t$ the phase has changed by

$$x \sum_{n=1}^{N} G_x(n\Delta)\Delta,$$

which is approximately $x \int_0^t G_x(s)ds$.

20.6.1 An Example of $\mathbf{G}(t)$

Suppose now that $g > 0$ and θ is an arbitrary direction vector. Let

$$\mathbf{G}(t) = g\theta, \text{ for } \tau \leq t, \tag{20.3}$$

and $\mathbf{G}(t) = 0$ otherwise. Then the received signal $S(t)$ is

$$S(t) = \int_{R^3} \rho(\mathbf{r}) \exp(-i\gamma g(t-\tau)\theta \cdot \mathbf{r}) d\mathbf{r}$$

$$= (2\pi)^{3/2} \hat{\rho}(\gamma g(t-\tau)\theta), \tag{20.4}$$

for $\tau \leq t << T_2$, where $\hat{\rho}$ denotes the three-dimensional Fourier transform of the function $\rho(\mathbf{r})$.

From Equation (20.4) we see that, by selecting different direction vectors and by sampling the received signal $S(t)$ at various times, we can obtain values of the Fourier transform of ρ along lines through the origin in the Fourier domain, called *k-space*. If we had these values for all θ and for all t we would be able to determine $\rho(\mathbf{r})$ exactly. Instead, we have much the same problem as in transmission tomography; only finitely many θ and only finitely many samples of $S(t)$. Noise is also a problem, because the resonance signal is not strong, even though the external magnetic field is.

We may wish to avoid having to estimate the function $\rho(\mathbf{r})$ from finitely many noisy values of its Fourier transform. We can do this by selecting the gradient field $\mathbf{G}(t)$ differently.

20.6.2 Another Example of $\mathbf{G}(t)$

The vector-valued function $\mathbf{G}(t)$ can be written as

$$\mathbf{G}(t) = (G_1(t), G_2(t), G_3(t)).$$

Now we let

$$G_2(t) = g_2,$$

and

$$G_3(t) = g_3,$$

for $0 \leq t \leq \tau$, and zero otherwise, and

$$G_1(t) = g_1,$$

for $\tau \leq t$, and zero otherwise. This means that only $H_0\mathbf{k}$ and the rf field are present up to time τ, and then the rf field is shut off and the gradient field is turned on. Then, for $t \geq \tau$, we have

$$S(t) = (2\pi)^{3/2} \hat{M}_0(\gamma(t-\tau)g_1, \gamma\tau g_2, \gamma\tau g_3).$$

By selecting

$$t_n = n\Delta t + \tau, \text{for } n = 1, ..., N,$$

$$g_{2k} = k\Delta g,$$

and

$$g_{3i} = i\Delta g,$$

for $i, k = -m, ..., m$ we have values of the Fourier transform, $\hat{M}_0$, on a Cartesian grid in three-dimensional k-space. The proton density function, ρ, can then be approximated using the fast Fourier transform.

Although the reconstruction employs the FFT, obtaining the Fourier-transform values on the Cartesian grid can take time. An abdominal scan can last for a couple of hours, during which the patient is confined, motionless and required to hold his or her breath repeatedly. Recent work on *compressed sensing* is being applied to reduce the number of Fourier-transform values that need to be collected, and thereby reduce the scan time [270, 199].

20.7 Compressed Sensing in Image Reconstruction

As we have seen, the data one obtains from the scanning process can often be interpreted as values of the Fourier transform of the desired image; this is precisely the case in magnetic-resonance imaging, and approximately true for x-ray transmission tomography, positron-emission tomography (PET) and single-photon emission tomography (SPECT). The images one encounters in medical diagnosis are often approximately locally constant, so the associated array of discrete partial derivatives will be sparse. If this sparse derivative array can be recovered from relatively few Fourier-transform values, then the scanning time can be reduced.

20.7.1 Incoherent Bases

The objective in CS is to exploit sparseness to reconstruct a vector f in $\mathbb{R}^J$ from relatively few linear functional measurements [116].

Let $U = \{u^1, u^2, ..., u^J\}$ and $V = \{v^1, v^2, ..., v^J\}$ be two orthonormal bases for $\mathbb{R}^J$, with all members of $\mathbb{R}^J$ represented as column vectors. For $i = 1, 2, ..., J$, let

$$\mu_i = \max_{1 \leq j \leq J} \{|\langle u^i, v^j \rangle|\}$$

and

$$\mu(U, V) = \max\{mu_i \,|i = 1, ..., I\}.$$

We know from Cauchy's Inequality that

$$|\langle u^i, v^j \rangle| \leq 1,$$

and from Parseval's Equation

$$\sum_{j=1}^{J} |\langle u^i, v^j \rangle|^2 = ||u^i||^2 = 1.$$

Therefore, we have

$$\frac{1}{\sqrt{J}} \leq \mu(U, V) \leq 1.$$

The quantity $\mu(U, V)$ is the *coherence* measure of the two bases; the closer $\mu(U, V)$ is to the lower bound of $\frac{1}{\sqrt{J}}$, the more *incoherent* the two bases are.

Let f be a fixed member of $\mathbb{R}^J$; we expand f in the V basis as

$$f = x_1 v^1 + x_2 v^2 + ... + x_J v^J.$$

We say that the coefficient vector $x = (x_1, ..., x_J)$ is S-sparse if S is the number of non-zero x_j.

20.7.2 Exploiting Sparseness

If S is small, most of the x_j are zero, but since we do not know which ones these are, we would have to compute all the linear functional values

$$x_j = \langle f, v^j \rangle$$

to recover f exactly. In fact, the smaller S is, the harder it would be to learn anything from randomly selected x_j, since most would be zero. The idea in CS is to obtain measurements of f with members of a different orthonormal basis, which we call the U basis. If the members of U are very much like the members of V, then nothing is gained. But, if the members of U are quite unlike the members of V, then each inner product measurement

$$y_i = \langle f, u^i \rangle = f^T u^i$$

should tell us something about f. If the two bases are sufficiently incoherent, then relatively few y_i values should tell us quite a bit about f. Specifically, we have the following result due to Candès and Romberg [72]: suppose the coefficient vector x for representing f in the V basis is S-sparse. Select uniformly randomly $M \leq J$ members of the U basis and compute the measurements $y_i = \langle f, u^i \rangle$. Then, if M is sufficiently large, it is highly probable that $z = x$ also solves the problem of minimizing the one-norm

$$||z||_1 = |z_1| + |z_2| + ... + |z_J|,$$

subject to the conditions

$$y_i = \langle g, a^i \rangle = g^T u^i,$$

for those M randomly selected u^i, where

$$g = z_1 v^1 + z_2 v^2 + ... + z_J v^J.$$

This can be formulated as a linear programming problem. The smaller $\mu(U, V)$ is, the smaller the M is permitted to be without reducing the probability of perfect reconstruction.

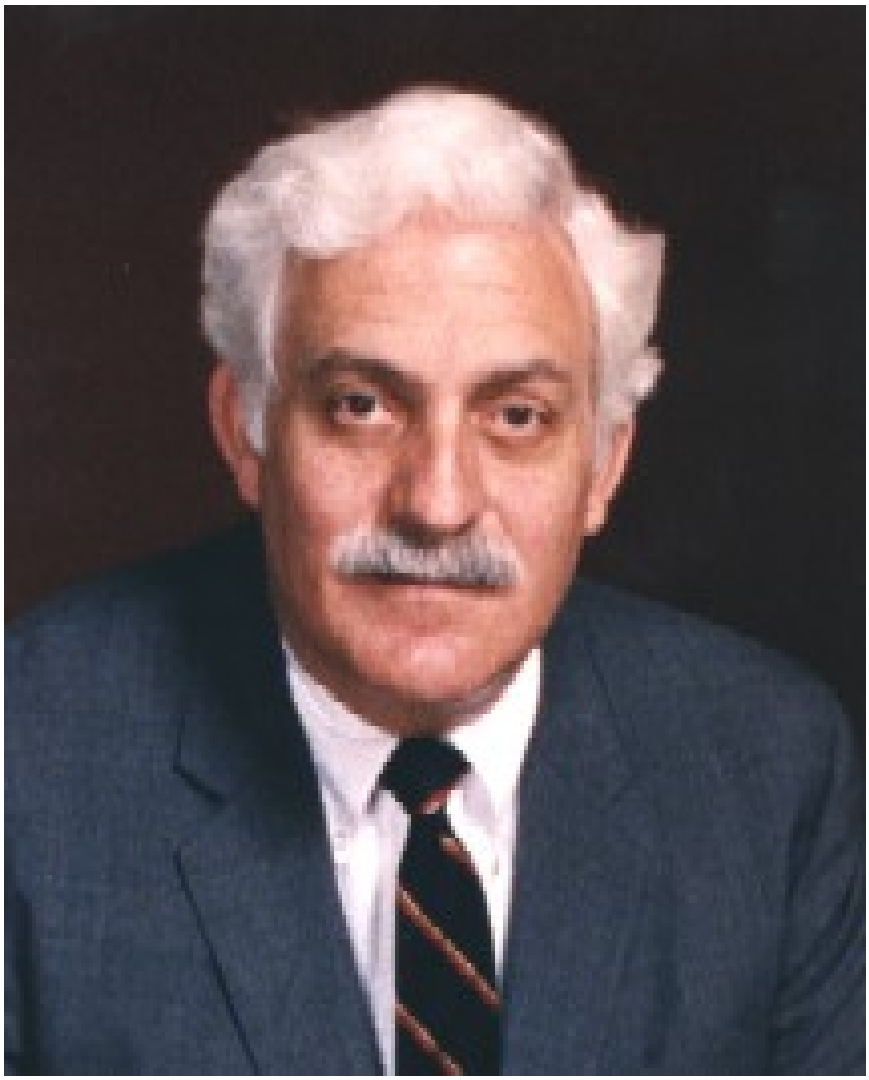

FIGURE 20.1: Raymond Damadian: inventor of MRI.

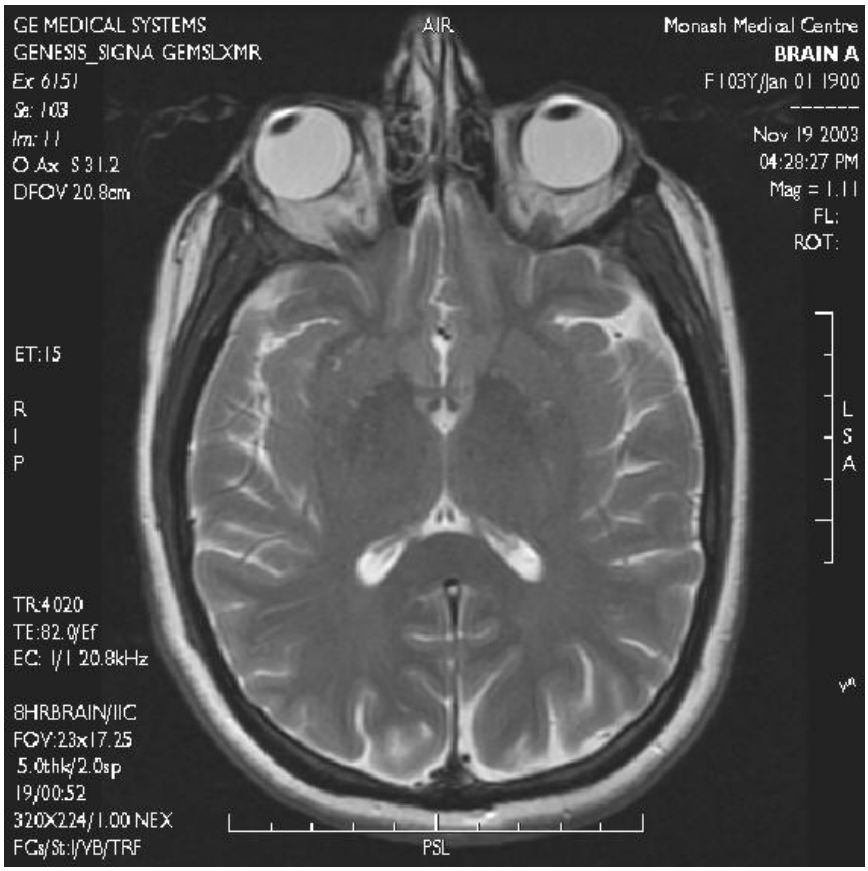

FIGURE 20.2: An MRI head scan. Check out the eyeballs.

Chapter 21

Intensity Modulated Radiation Therapy

21.1 Chapter Summary

In *intensity modulated radiation therapy* (IMRT) beamlets of radiation with different intensities are transmitted into the body of the patient. Each voxel within the patient will then absorb a certain dose of radiation from each beamlet. The goal of IMRT is to direct a sufficient dosage to those regions requiring the radiation, those that are designated *planned target volumes* (PTV), while limiting the dosage received by the other regions, the so-called *organs at risk* (OAR). In our discussion here we follow Censor et al. [82].

21.2 The Forward and Inverse Problems

The *forward problem* is to calculate the radiation dose absorbed in the irradiated tissue based on a given distribution of the beamlet intensities. The *inverse problem* is to find a distribution of beamlet intensities, the radiation intensity map, that will result in a clinically acceptable dose distribution. One important constraint is that the radiation intensity map must be implementable, that is, it is physically possible to produce such

an intensity map, given the machine's design. There will be limits on the change in intensity between two adjacent beamlets, for example.

21.3 Equivalent Uniform Dosage

The *equivalent uniform dose* (EUD) for tumors is the biologically equivalent dose which, if given uniformly, will lead to the same cell-kill within the tumor volume as the actual non-uniform dose.

21.4 Constraints

Constraints on the EUD received by each voxel of the body are described in *dose space*, the space of vectors whose entries are the doses received at each voxel. Constraints on the deliverable radiation intensities of the beamlets are best described in *intensity space*, the space of vectors whose entries are the intensity levels associated with each of the beamlets. The constraints in dose space will be upper bounds on the dosage received by the OAR and lower bounds on the dosage received by the PTV. The constraints in intensity space are limits on the complexity of the intensity map and on the delivery time, and, obviously, that the intensities be non-negative. Because the constraints operate in two different domains, it is convenient to formulate the problem using these two domains. This leads to a split-feasibility problem.

21.5 The Multi-Set Split-Feasibilty-Problem Model

The *split feasibility problem* (SFP) is to find an x in a given closed convex subset C of $\mathbb{R}^J$ such that Ax is in a given closed convex subset Q of $\mathbb{R}^I$, where A is a given real I by J matrix. Because the constraints are best described in terms of several sets in dose space and several sets in intensity space, the SFP model needed to be expanded into the *multi-set* SFP (MSSFP) [84] and the CQ algorithm extended to this more general case.

It is not uncommon to find that, once the various constraints have been

specified, there is no intensity map that satisfies them all. In such cases, it is desirable to find an intensity map that comes as close as possible to satisfying all the constraints. One way to do this, as we shall see, is to minimize a *proximity function*.

21.6 Formulating the Proximity Function

For $i = 1, ..., I$, and $j = 1, ..., J$, let $h_i \geq 0$ be the dose absorbed by the i-th voxel of the patient's body, $x_j \geq 0$ be the intensity of the j-th beamlet of radiation, and $D_{ij} \geq 0$ be the dose absorbed at the i-th voxel due to a unit intensity of radiation at the j-th beamlet. The non-negative matrix D with entries D_{ij} is the *dose influence matrix*.

In intensity space, we have the obvious constraints that $x_j \geq 0$. In addition, there are *implementation constraints*; the available treatment machine will impose its own requirements, such as a limit on the difference in intensities between adjacent beamlets. In dosage space, there will be a lower bound on the dosage delivered to those regions designated as *planned target volumes* (PTV), and an upper bound on the dosage delivered to those regions designated as *organs at risk* (OAR).

21.7 Equivalent Uniform Dosage Functions

Suppose that S_t is either a PTV or a OAR, and suppose that S_t contains N_t voxels. For each dosage vector $h = (h_1, ..., h_I)^T$ define the *equivalent uniform dosage function* (EUD-function) $e_t(h)$ by

$$e_t(h) = (\frac{1}{N_t} \sum_{i \in S_t} (h_i)^\alpha)^{1/\alpha}, \tag{21.1}$$

where $0 < \alpha < 1$ if S_t is a PTV, and $\alpha > 1$ if S_t is an OAR. The function $e_t(h)$ is convex, for h nonnegative, when S_t is an OAR, and $-e_t(h)$ is convex, when S_t is a PTV. The constraints in dosage space take the form

$$e_t(h) \leq a_t,$$

when S_t is an OAR, and

$$-e_t(h) \leq b_t,$$

when S_t is a PTV. Therefore, we require that $h = Dx$ lie within the intersection of these convex sets.

21.8 Recent Developments

One drawback to the use of x-rays in radiation therapy is that they continue through the body after they have encountered their target. A recent technology, proton-beam therapy, directs a beam of protons at the target. Since the protons are heavy, and have mass and charge, their trajectories can be controlled in ways that x-ray trajectories cannot be. The new proton center at Massachusetts General Hospital in Boston is one of the first to have this latest technology. As with most new and expensive medical procedures, there is some debate going on about just how much of an improvement it provides, relative to other methods.

Part V

Appendices

Chapter 22

Appendix: Linear Algebra

22.1 Chapter Summary

Linear algebra is the study of linear transformations between vector spaces. Although the subject is not simply matrix theory, there is a close connection, stemming from the role of matrices in representing linear transformations. Throughout this section we shall limit discussion to finite-dimensional vector spaces.

22.2 Representing a Linear Transformation

Let $\mathcal{A} = \{a^1, a^2, ..., a^N\}$ be a basis for the finite-dimensional complex vector space V. Now that the basis for V is specified, there is a natural association, an *isomorphism*, between V and the vector space $\mathbb{C}^N$ of N-

dimensional column vectors with complex entries. Any vector v in V can be written as

$$v = \sum_{n=1}^{N} \gamma_n a^n. \tag{22.1}$$

The column vector $\gamma = (\gamma_1, ..., \gamma_N)^T$ is uniquely determined by v and the basis $\mathcal{A}$ and we denote it by $[v]_{\mathcal{A}}$. Notice that the ordering of the list of members of $\mathcal{A}$ matters, so we shall always assume that the ordering has been fixed.

Let W be a second finite-dimensional vector space, and let T be any linear transformation from V to W. Let $\mathcal{B} = \{b^1, b^2, ..., b^M\}$ be a basis for W. For $n = 1, ..., N$, let

$$Ta^n = A_{1n}b^1 + A_{2n}b^2 + ... + A_{Mn}b^M. \tag{22.2}$$

Then the M by N matrix A having the A_{mn} as entries is said to *represent* T, with respect to the bases $\mathcal{A}$ and $\mathcal{B}$, and we write $A = [T]_{\mathcal{A}}^{\mathcal{B}}$.

Ex. 22.1 *Show that* $[Tv]_{\mathcal{B}} = A[v]_{\mathcal{A}}$.

Ex. 22.2 *Let* P_2 *and* P_3 *be the vector spaces of real polynomials of degrees two and three, respectively. Let* $T : P_3 \to P_2$ *be the differentiation operator. Select bases for* P_2 *and* P_3 *and represent* T *by matrix multiplication.*

Ex. 22.3 *Suppose that* V, W *and* Z *are vector spaces, with bases* $\mathcal{A}$, $\mathcal{B}$ *and* $\mathcal{C}$, *respectively. Suppose also that* T *is a linear transformation from* V *to* W *and* U *is a linear transformation from* W *to* Z. *Let* A *represent* T *with respect to the bases* $\mathcal{A}$ *and* $\mathcal{B}$, *and let* B *represent* U *with respect to the bases* $\mathcal{B}$ *and* $\mathcal{C}$. *Show that the matrix* BA *represents the linear transformation* UT *with respect to the bases* $\mathcal{A}$ *and* $\mathcal{C}$.

22.3 Linear Operators on V

When $W = V$, we say that the linear transformation T is a *linear operator* on V. In this case, we can also take the basis $\mathcal{B}$ to be $\mathcal{A}$, and say that the matrix A represents the linear operator T, with respect to the basis $\mathcal{A}$. We then write $A = [T]_{\mathcal{A}}$.

Ex. 22.4 *Suppose that* $\tilde{\mathcal{A}}$ *is a second basis for* V. *Let* T *be any linear operator on* V *and* $\tilde{A} = [T]_{\tilde{\mathcal{A}}}$. *Show that there is a unique invertible* N *by*

N matrix Q having the property that, for all T, the matrix $\tilde{A} = QAQ^{-1}$, so we can write

$$[T]_{\tilde{\mathcal{A}}} = Q[T]_{\mathcal{A}}Q^{-1}.$$

Hint: the matrix Q is the change-of-basis matrix*, which means that Q represents the identity operator I, with respect to the bases $\mathcal{A}$ and $\tilde{\mathcal{A}}$; that is, $Q = [I]^{\tilde{\mathcal{A}}}_{\mathcal{A}}$.*

Ex. 22.5 *Let T be a linear operator on the finite-dimensional vector space V with basis $\mathcal{A} = \{a^1, a^2, ..., a^N\}$. Let W be the subspace of V spanned by the elements $\{a^1, ..., a^M\}$, where $M < N$. Suppose that W is T-invariant, that is, $Tw \in W$ for every $w \in W$. What can then be said about the representing matrix $A = [T]_{\mathcal{A}}$?*

22.4 Linear Operators on $\mathbb{C}^N$

Let $\mathcal{A}$ be the usual basis for the vector space $V = \mathbb{C}^N$. In practice, we make no distinction between a member x of $\mathbb{C}^N$ and $[x]_{\mathcal{A}}$; that is, we use the equation

$$x = [x]_{\mathcal{A}}$$

without comment. If T is a linear operator on $\mathbb{C}^N$ and $A = [T]_{\mathcal{A}}$, then from

$$[Tx]_{\mathcal{A}} = A[x]_{\mathcal{A}}$$

we write

$$Tx = Ax;$$

in other words, we make no distinction between T and A and say that every linear operator on $\mathbb{C}^N$ is multiplication by a matrix. Of course, all of this presupposes that $\mathcal{A}$ is the usual basis for $\mathbb{C}^N$; if we change the basis, then the distinctions again become necessary.

22.5 Similarity and Equivalence of Matrices

Let $\mathcal{A}$ and $\tilde{\mathcal{A}} = \{\tilde{a}^1, ..., \tilde{a}^N\}$ be bases for V, and $\mathcal{B}$ and $\tilde{\mathcal{B}} = \{\tilde{b}^1, ..., \tilde{b}^M\}$ be bases for W. Let $Q = [I]^{\tilde{\mathcal{A}}}_{\mathcal{A}}$ and $R = [I]^{\tilde{\mathcal{B}}}_{\mathcal{B}}$ be the change-of-bases matrices

in V and W, respectively. As we just saw, for any linear operator T on V, the matrices $\tilde{A} = [T]_{\tilde{\mathcal{A}}}$ and $A = [T]_{\mathcal{A}}$ are related according to

$$A = Q^{-1}\tilde{A}Q. \tag{22.3}$$

We describe the relationship in Equation (22.3) by saying that the matrices A and $\tilde{A}$ are *similar*.

Definition 22.1 *Two N by N matrices A and B are said to be* similar *if there is an invertible matrix Q such that $A = Q^{-1}BQ$.*

Ex. 22.6 *Show that similar matrices have the same eigenvalues.*

Let S be a linear transformation from V to W. Then we have

$$[S]_{\mathcal{A}}^{\mathcal{B}} = R^{-1}[S]_{\tilde{\mathcal{A}}}^{\tilde{\mathcal{B}}}Q. \tag{22.4}$$

With $G = [S]_{\mathcal{A}}^{\mathcal{B}}$ and $\tilde{G} = [S]_{\tilde{\mathcal{A}}}^{\tilde{\mathcal{B}}}$, we have

$$G = R^{-1}\tilde{G}Q. \tag{22.5}$$

Definition 22.2 *Two M by N matrices A and B are said to be* equivalent *if there are invertible matrices P and Q such that $B = PAQ$.*

We can therefore describe the relationship in Equation (22.5) by saying that the matrices G and $\tilde{G}$ are equivalent.

Ex. 22.7 *Show that A and B are equivalent if B can be obtained from A by means of elementary row and column operations.*

Ex. 22.8 *Prove that two equivalent matrices A and B must have the same rank, and so two similar matrices must also have the same rank. Hint: use the fact that Q is invertible to show that A and AQ have the same rank.*

Ex. 22.9 *Prove that any two M by N matrices with the same rank r are equivalent. Hints: Let A be an M by N matrix, which we can also view as inducing, by multiplication, a linear transformation T from $V = \mathbb{C}^N$ to $W = \mathbb{C}^M$. Therefore, A represents T in the usual bases of $\mathbb{C}^N$ and $\mathbb{C}^M$. Now construct a basis $\mathcal{A}$ for $\mathbb{C}^N$, such that*

$$\mathcal{A} = \{a^1, ..., a^N\},$$

with $\{a^{r+1}, ..., a^N\}$ forming a basis for the null space of A. Show that the set $\{Aa^1, ..., Aa^r\}$ is linearly independent and can therefore be extended to a basis $\mathcal{B}$ for $\mathbb{C}^M$. Show that the matrix D that represents T with respect to the bases $\mathcal{A}$ and $\mathcal{B}$ is the M by N matrix with the r by r identity matrix in the upper left corner, and all the other entries are zero. Since A is then equivalent to this matrix D, so is the matrix B; therefore A and B are equivalent to each other. Another way to say this is that both A and B can be reduced to D using elementary row and column operations.

22.6 Linear Functionals and Duality

We turn now to the particular case in which the second vector space W is just the space $\mathbb{C}$ of complex numbers. Any linear transformation f from V to $\mathbb{C}$ is called a *linear functional*. The space of all linear functionals on V is denoted V^* and called the *dual space* of V. The set V^* is itself a finite-dimensional vector space, so it too has a dual space, $(V^*)^* = V^{**}$, the second dual space, which is the set of all linear transformations F from V^* to $\mathbb{C}$.

Ex. 22.10 *Show that the dimension of V^* is the same as that of V. Hint: let $\mathcal{A} = \{a^1, ..., a^N\}$ be a basis for V, and for each $m = 1, ..., N$, let $f^m(a^n) = 0$, if $m \neq n$, and $f^m(a^m) = 1$. Show that the collection $\{f^1, ..., f^N\}$ is a basis for V^*.*

Proposition 22.1 *Let V be a vector space of dimension N and S a subspace of V. Then the dimension of S is $N - 1$ if and only if there is a non-zero member f of V^* such that $S = \{v | f(v) = 0\}$.*

Proof: Let S have dimension $M < N$ and let $\{u^1, u^2, ..., u^M\}$ be a basis for S. Extend this basis for S to a basis for V, denoted

$$\{u^1, u^2, ..., u^M, v^1, v^2, ..., v^{N-M}\}.$$

Now suppose that the dimension of S is $M = N - 1$, and that the enlarged basis has only one new member, v^1. Every vector v in V can be written uniquely as

$$v = a_1 u^1 + a_2 u^2 + ... + a_{N-1} u^{N-1} + a_N v^1.$$

Let $f(v) = a_N$; then f is a member of V^* and $S = \{v | f(v) = 0\}$.

Conversely, suppose now that $S = \{v | f(v) = 0\}$, and its dimension is $M < N - 1$. Then the enlarged basis has at least two new members, v^1 and v^2, neither of them in S. Therefore $\alpha_1 = f(v^1)$ and $\alpha_2 = f(v^2)$ are not zero. We then have $f(v) = 0$ for the vector $v = \alpha_2 v^1 - \alpha_1 v^2$, which means that v is in S. But v is a linear combination of v^1 and v^2, and therefore, because of the linear independence of the members of the enlarged basis, cannot also be a linear combination of the u^m, for $m = 1, 2, ..., M$. ∎

There is a natural identification of V^{**} with V itself. For each v in V, define $J_v(f) = f(v)$ for each f in V^*. Then it is easy to establish that J_v is in V^{**} for each v in V. The set J_V of all members of V^{**} of the form J_v for some v is a subspace of V^{**}.

Ex. 22.11 *Show that the subspace J_V has the same dimension as V^{**} itself, so that it must be all of V^{**}.*

In the previous exercise we established that $J_V = V^{**}$ by showing that these spaces have the same dimension. We can also prove this result in a more direct way. Let F be any member of V^{**}. We show that there is a v in V such that $F(f) = f(v)$ for all f in V^* by displaying v explicitly. Let $\gamma_n = F(f^n)$, for $n = 1, 2, ..., N$, where f^n are as defined in Exercise 22.10. Then let $v = \gamma_1 a^1 + \gamma_2 a^2 + ... + \gamma_N a^N$. Let f be arbitrary in V^*, written in terms of the basis as

$$f = \alpha_1 f^1 + \alpha_2 f^2 + ... + \alpha_N f^N,$$

so that

$$f(v) = \alpha_1 f^1(v) + \alpha_2 f^2(v) + ... + \alpha_N f^N(v) = \alpha_1\gamma_1 + \alpha_2\gamma_2 + ... + \alpha_N\gamma_N.$$

Then

$$F(f) = \alpha_1 F(f^1) + \alpha_2 F(f^2) + ... + \alpha_N F(f^N) = \alpha_1\gamma_1 + \alpha_2\gamma_2 + ... + \alpha_N\gamma_N = f(v).$$

We shall see later that once V has been endowed with an inner product, there is a simple way to describe every linear functional on V: for each f in V^* there is a unique vector v_f in V with $f(v) = \langle v, v_f \rangle$, for each v in V. As a result, we have an identification of V^* with V itself.

22.7 Diagonalization

Let $T : V \to V$ be a linear operator, $\mathcal{A}$ a basis for V, and $A = [T]_{\mathcal{A}}$. As we change the basis, the matrix representing T also changes. We wonder if it is possible to find some basis $\mathcal{B}$ such that $B = [T]_{\mathcal{B}}$ is a diagonal matrix L. Let $P = [I]_{\mathcal{B}}^{\mathcal{A}}$ be the change-of basis matrix from $\mathcal{B}$ to $\mathcal{A}$. We would then have $P^{-1}AP = L$, or $A = PLP^{-1}$. When this happens, we say that A has been *diagonalized* by P. According to Lemma 6.6, A is diagonalizable if all its eigenvalues are distinct.

Suppose that the basis $\mathcal{B} = \{b^1, ..., b^N\}$ is such that $B = [T]_{\mathcal{B}} = L$, where L is the diagonal matrix $L = \text{diag}\{\lambda_1, ..., \lambda_N\}$. Then we have $AP = PL$, which tells us that p^n, the n-th column of P, is an eigenvector of the matrix A, with λ_n as its eigenvalue. Since $p^n = [b^n]_{\mathcal{A}}$, we have

$$0 = (A - \lambda_n I)p^n = (A - \lambda_n I)[b^n]_{\mathcal{A}} = [(T - \lambda_n I)b^n]_{\mathcal{A}},$$

from which we conclude that

$$(T - \lambda_n I)b^n = 0,$$

or

$$Tb^n = \lambda_n b^n;$$

therefore, b^n is an eigenvector of the linear operator T.

22.8 Using Matrix Representations

The matrix A has eigenvalues λ_n, $n = 1, ..., N$, precisely when these λ_n are the roots of the *characteristic polynomial*

$$P(\lambda) = \det(A - \lambda I).$$

We would like to be able to define the characteristic polynomial of T itself to be $P(\lambda)$; the problem is that we do not yet know that different matrix representations of T have the same characteristic polynomial, although we do know that, since they are similar matrices, they have the same eigenvalues.

Ex. 22.12 *Use the fact that det(GH)=det(G)det(H) for any square matrices G and H to show that*

$$\det([T]_{\mathcal{B}} - \lambda I) = \det([T]_{\mathcal{C}} - \lambda I),$$

for any bases $\mathcal{B}$ and $\mathcal{C}$ for V.

22.9 An Inner Product on V

For any two column vectors $x = (x_1, ..., x_N)^T$ and $y = (y_1, ..., y_n)^T$ in $\mathbb{C}^N$, their *complex dot product* is defined by

$$x \cdot y = \sum_{n=1}^{N} x_n \overline{y_n} = y^\dagger x,$$

where $y^\dagger$ is the *conjugate transpose* of the vector y, that is, $y^\dagger$ is the row vector with entries $\overline{y_n}$.

The association of the elements v in V with the complex column vector $[v]_{\mathcal{A}}$ can be used to obtain an *inner product* on V. For any v and w in V, define

$$\langle v, w \rangle = [v]_{\mathcal{A}} \cdot [w]_{\mathcal{A}}, \tag{22.6}$$

where the right side is the ordinary complex dot product in $\mathbb{C}^N$. Note that, with respect to this inner product, the basis $\mathcal{A}$ becomes an orthonormal basis.

For particular vector spaces V we may want to define an inner product that conforms well to the special nature of the elements of V. For example, suppose that V is the vector space of all N by N complex matrices. This space has dimension N^2. A basis for this space is the collection of all N by N matrices that have a one in a single entry and zero everywhere else. The induced inner product that we get using this basis can be described in another way: it is $\langle A, B\rangle = \text{trace}(B^\dagger A)$. The resulting norm of A is the *Frobenius norm*.

22.10 Orthogonality

Two vectors v and w in the inner-product space V are said to be *orthogonal* if $\langle v, w\rangle = 0$. A basis $\mathcal{U} = \{u^1, u^2, ..., u^N\}$ is called an *orthogonal basis* if every two vectors in $\mathcal{U}$ are orthogonal, and *orthonormal* if, in addition, $\|u^n\| = 1$, for each n.

Ex. 22.13 *Let $\mathcal{U}$ and $\mathcal{V}$ be orthonormal bases for the inner-product space V, and let Q be the change-of-basis matrix satisfying*

$$[v]_\mathcal{U} = Q[v]_\mathcal{V}.$$

Show that $Q^{-1} = Q^\dagger$, so that Q is a unitary matrix.

22.11 Representing Linear Functionals

Let $f : V \to \mathbb{C}$ be a linear functional on the inner-product space V and let $\mathcal{A} = \{a^1, ..., a^N\}$ be an orthonormal basis for V. Let v_f be the member of V defined by

$$v_f = \sum_{m=1}^{N} \overline{f(a^m)} a^m.$$

Then for each

$$v = \sum_{n=1}^{N} \alpha_n a^n,$$

in V, we have

$$\langle v, v_f\rangle = \sum_{n=1}^{N} \sum_{m=1}^{N} \alpha_n f(a^m) \langle a^n, a^m\rangle$$

$$= \sum_{n=1}^{N} \alpha_n f(a^n) = f(\sum_{n=1}^{N} \alpha_n a^n) = f(v).$$

So we see that once V has been given an inner product, each linear functional f on V can be thought of as corresponding to a vector v_f in V, so that

$$f(v) = \langle v, v_f \rangle.$$

Ex. 22.14 *Show that the vector v_f associated with the linear functional f is unique by showing that*

$$\langle v, y \rangle = \langle v, w \rangle,$$

for every v in V implies that $y = w$.

22.12 Adjoint of a Linear Transformation

If T is a linear operator on an inner product space V, we say that T is *self adjoint* if $\langle Tu, v \rangle = \langle u, Tv \rangle$, for all u and v in V. This definition allows us to speak of self-adjoint linear operators before we have introduced the adjoint of a linear operator, the topic of this section.

Let $T : V \to W$ be a linear transformation from a vector space V to a vector space W. The *adjoint* of T is the linear operator $T^* : W^* \to V^*$ defined by

$$(T^* g)(v) = g(Tv), \tag{22.7}$$

for each $g \in W^*$ and $v \in V$.

Once V and W have been given inner products, and V^* and W^* have been identified with V and W, respectively, the operator T^* can be defined as a linear operator from W to V as follows. Let $T : V \to W$ be a linear transformation from an inner-product space V to an inner-product space W. For each fixed w in W, define a linear functional f on V by

$$f(v) = \langle Tv, w \rangle.$$

By our earlier discussion, f has an associated vector v_f in V such that

$$f(v) = \langle v, v_f \rangle.$$

Therefore,

$$\langle Tv, w \rangle = \langle v, v_f \rangle,$$

for each v in V. The *adjoint* of T is the linear transformation T^* from W to V defined by $T^*w = v_f$.

When $W = V$, and T is a linear operator on V, then so is T^*. In this case, we can ask whether or not $T^*T = TT^*$, that is, whether or not T is *normal*, and whether or not $T = T^*$, that is, whether or not T is *self-adjoint*.

Ex. 22.15 *Let $\mathcal{U}$ be an orthonormal basis for the inner-product space V and T a linear operator on V. Show that*

$$[T^*]_{\mathcal{U}} = ([T]_{\mathcal{U}})^\dagger. \tag{22.8}$$

22.13 Normal and Self-Adjoint Operators

Let T be a linear operator on an inner-product space V. We say that T is *normal* if $T^*T = TT^*$, and *self-adjoint* if $T^* = T$. A square matrix A is said to be *normal* if $A^\dagger A = AA^\dagger$, and *Hermitian* if $A^\dagger = A$.

Ex. 22.16 *Let $\mathcal{U}$ be an orthonormal basis for the inner-product space V. Show that T is normal if and only if $[T]_{\mathcal{U}}$ is a normal matrix, and T is self-adjoint if and only if $[T]_{\mathcal{U}}$ is Hermitian. Hint: use Exercise (22.3).*

Ex. 22.17 *Compute the eigenvalues for the real square matrix*

$$A = \begin{bmatrix} 1 & 2 \\ -2 & 1 \end{bmatrix}. \tag{22.9}$$

Note that the eigenvalues are complex, even though the entries of A are real. The matrix A is not Hermitian.

Ex. 22.18 *Show that the eigenvalues of the complex matrix*

$$B = \begin{bmatrix} 1 & 2+i \\ 2-i & 1 \end{bmatrix} \tag{22.10}$$

are the real numbers $\lambda = 1 + \sqrt{5}$ and $\lambda = 1 - \sqrt{5}$, with corresponding eigenvectors $u = (\sqrt{5}, 2-i)^T$ and $v = (\sqrt{5}, i-2)^T$, respectively.

Ex. 22.19 *Show that the eigenvalues of the real matrix*

$$C = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \tag{22.11}$$

are both equal to one, and that the only eigenvectors are non-zero multiples of the vector $(1, 0)^T$. Compute C^TC and CC^T. Are they equal?

22.14 It is Good to be "Normal"

For a given linear operator on V, when does there exist an orthonormal basis for V consisting of eigenvectors of T? The answer is: When T is normal.

Consider an N by N matrix A. We use A to define a linear operator T on the space of column vectors $V = \mathbb{C}^N$ by $Tv = Av$, that is, the operator T works by multiplying each column vector v in $\mathbb{C}^N$ by the matrix A. Then A represents T with respect to the usual orthonormal basis $\mathcal{A}$ for $\mathbb{C}^N$. Suppose now that there is an orthonormal basis $\mathcal{U} = \{u^1, ..., u^N\}$ for $\mathbb{C}^N$ such that

$$Tu^n = Au^n = \lambda_n u^n,$$

for each n. The matrix representing T in the basis $\mathcal{U}$ is the matrix $B = Q^{-1}AQ$, where Q is the change-of-basis matrix with

$$Q[v]_{\mathcal{U}} = [v]_{\mathcal{A}}.$$

But we also know that B is the diagonal matrix $B = L = \text{diag}(\lambda_1, ..., \lambda_N)$. Therefore, $L = Q^{-1}AQ$, or $A = QLQ^{-1}$.

As we saw in Exercise (22.13), the matrix Q is unitary, that is, $Q^{-1} = Q^\dagger$. Therefore, $A = QLQ^\dagger$. Then we have

$$A^\dagger A = QL^\dagger Q^\dagger QLQ^\dagger = QL^\dagger LQ^\dagger$$

$$= QLL^\dagger Q^\dagger = QLQ^\dagger QL^\dagger Q^\dagger = AA^\dagger,$$

so that

$$A^\dagger A = AA^\dagger,$$

and A is normal.

Two fundamental results in linear algebra are the following, which we discuss in more detail in the chapter "Hermitian and Normal Linear Operators".

Theorem 22.1 *For a linear operator T on a finite-dimensional complex inner-product space V there is an orthonormal basis of eigenvectors if and only if T is normal.*

Corollary 22.1 *A self-adjoint linear operator T on a finite-dimensional complex inner-product space V has an orthonormal basis of eigenvectors.*

Ex. 22.20 *Show that the eigenvalues of a self-adjoint linear operator T on a finite-dimensional complex inner-product space are real numbers. Hint: consider $Tu = \lambda u$, and begin with $\lambda\langle u, u\rangle = \langle Tu, u\rangle$.*

Combining the various results obtained so far, we can conclude the following.

Corollary 22.2 *Let T be a linear operator on a finite-dimensional real inner-product space V. Then V has an orthonormal basis consisting of eigenvectors of T if and only if T is self-adjoint.*

22.15 Bases and Inner Products

Throughout this section V will denote a finite-dimensional real or complex vector space. We know that it is always possible to find a basis for V; we simply build up a set of linearly independent vectors until including any additional vector will render the set linearly dependent. As we have seen, once we have a basis for V it is a simple matter to use that basis to induce an inner product on V. In this section we make several assertions without proof; the proofs are left as exercises for the reader.

Let $\mathcal{A} = \{a^1, ..., a^N\}$ be a basis for V. Each vector x in V can then be written uniquely as a linear combination of the members of $\mathcal{A}$:

$$x = \alpha_1 a^1 + ... + \alpha_N a^N.$$

The column vector $\alpha = (\alpha_1, ..., \alpha_N)^T$ is then denoted $[x]_{\mathcal{A}}$. We denote by $F_{\mathcal{A}}$ the linear transformation $F_{\mathcal{A}} : V \to \mathbb{C}^N$ that associates with each x in V the column vector $[x]_{\mathcal{A}}$, and by $E_{\mathcal{A}}$ the linear transformation $E_{\mathcal{A}} : \mathbb{C}^N \to V$ that associates with each vector α in $\mathbb{C}^N$ the member of V given by

$$x = \alpha_1 a^1 + ... + \alpha_N a^N.$$

Note that $E_{\mathcal{A}}$ is the inverse of $F_{\mathcal{A}}$.

The inner product on V induced by the basis $\mathcal{A}$ is

$$\langle x, y \rangle_{\mathcal{A}} = [x]_{\mathcal{A}} \cdot [y]_{\mathcal{A}},$$

which can also be written as

$$\langle x, y \rangle_{\mathcal{A}} = F_{\mathcal{A}} x \cdot F_{\mathcal{A}} y.$$

The basis $\mathcal{A}$ is orthonormal with respect to this inner product. We denote by $V_{\mathcal{A}}$ the vector space V with the inner product $\langle x, y \rangle_{\mathcal{A}}$.

The adjoint of $F_{\mathcal{A}}$ is the linear transformation $F_{\mathcal{A}}^* : \mathbb{C}^N \to V_{\mathcal{A}}$ for which

$$\langle F_{\mathcal{A}}^* \alpha, y \rangle_{\mathcal{A}} = \alpha \cdot F_{\mathcal{A}} y,$$

for all α in $\mathbb{C}^N$ and y in V. But we also have

$$\langle F_{\mathcal{A}}^* \alpha, y\rangle_{\mathcal{A}} = F_{\mathcal{A}} F_{\mathcal{A}}^* \alpha \cdot F_{\mathcal{A}} y.$$

It follows that

$$F_{\mathcal{A}} F_{\mathcal{A}}^* = I.$$

Therefore,

$$F_{\mathcal{A}}^* = E_{\mathcal{A}}.$$

Let $\mathcal{B} = \{b^1, ..., b^N\}$ be a second basis for V. The change-of-basis matrix $Q = [I]_{\mathcal{A}}^{\mathcal{B}}$ has the property

$$[x]_{\mathcal{B}} = Q[x]_{\mathcal{A}},$$

or

$$F_{\mathcal{B}} x = Q F_{\mathcal{A}} x,$$

for all x in V. Therefore we can write

$$F_{\mathcal{B}} = Q F_{\mathcal{A}},$$

so that

$$Q = F_{\mathcal{B}} E_{\mathcal{A}}.$$

Ex. 22.21 *Viewing $F_{\mathcal{B}}$ as a linear transformation from the inner product space $V_{\mathcal{A}}$ to $\mathbb{C}^N$, show that the adjoint of $F_{\mathcal{B}}$ is the linear transformation $F'_{\mathcal{B}}$ given by $F'_{\mathcal{B}} = E_{\mathcal{A}} Q^\dagger$.*

Then we have

$$\langle x, y\rangle_{\mathcal{B}} = F_{\mathcal{B}} x \cdot F_{\mathcal{B}} y = Q F_{\mathcal{A}} x \cdot Q F_{\mathcal{A}} y = Q^\dagger Q F_{\mathcal{A}} x \cdot F_{\mathcal{A}} y.$$

Writing

$$H = Q^\dagger Q = F_{\mathcal{A}} F'_{\mathcal{B}} F_{\mathcal{B}} E_{\mathcal{A}},$$

where $F'_{\mathcal{B}} = E_{\mathcal{A}} Q^\dagger$ is the adjoint of the linear transformation $F_{\mathcal{B}}$, with respect to the vector space $V_{\mathcal{A}}$, we have

$$\langle x, y\rangle_{\mathcal{B}} = H F_{\mathcal{A}} x \cdot F_{\mathcal{A}} y.$$

The matrix H is hermitian and positive-definite.

Now let S be the linear transformation on V for which $H = [S]_{\mathcal{A}}$. This means that

$$H F_{\mathcal{A}} x = F_{\mathcal{A}} S x,$$

for all x in V. Then we can get an explicit description of S;

$$S = E_{\mathcal{A}} H F_{\mathcal{A}} = E_{\mathcal{A}} Q^\dagger Q F_{\mathcal{A}}.$$

This tells us that for any other basis $\mathcal{B}$ the associated inner product can be expressed in terms of the inner product from $\mathcal{A}$ by

$$\langle x, y\rangle_{\mathcal{B}} = \langle Sx, y\rangle_{\mathcal{A}}.$$

The linear operator S is self-adjoint and positive-definite on the inner product space $V_{\mathcal{A}}$.

If T is any self-adjoint, positive-definite linear operator on $V_{\mathcal{A}}$ then T induces another inner product, denoted $\langle x, y\rangle_T$, by

$$\langle x, y\rangle_T = \langle Tx, y\rangle_{\mathcal{A}}.$$

We also know that $V_{\mathcal{A}}$ has an orthonormal basis $\{u^1, ..., u^N\}$ of eigenvectors of T, with $Tu^n = \lambda_n u^n$. Let $b^n = \frac{1}{\sqrt{\lambda_n}} u^n$. Then the family $\mathcal{B} = \{b^1, ..., b^N\}$ is another basis for V and

$$\langle x, y\rangle_T = \langle x, y\rangle_{\mathcal{B}}.$$

If we begin with a vector space V that already has an inner product $\langle x, y\rangle$, then

$$\langle x, y\rangle = \langle x, y\rangle_{\mathcal{A}},$$

for any orthonormal basis $\mathcal{A}$.

We can summarize our findings as follows:

- 1. Any inner product $\langle x, y\rangle$ on V is $\langle x, y\rangle_{\mathcal{A}}$, for any orthonormal basis $\mathcal{A}$;
- 2. Any basis $\mathcal{A}$ induces an inner product, $\langle x, y\rangle_{\mathcal{A}}$;
- 3. If $\mathcal{A}$ and $\mathcal{B}$ are any two bases for V, then

  $$\langle x, y\rangle_{\mathcal{B}} = \langle Sx, y\rangle_{\mathcal{A}},$$

  for some self-adjoint, positive definite linear operator S on $V_{\mathcal{A}}$;
- 4. If T is any self-adjoint positive-definite linear operator on $V_{\mathcal{A}}$, then T induces an inner product

  $$\langle x, y\rangle_T = \langle Tx, y\rangle,$$

  and there is a basis $\mathcal{B}$ such that

  $$\langle x, y\rangle_T = \langle x, y\rangle_{\mathcal{B}}.$$

Chapter 23

Appendix: Even More Matrix Theory

23.1 LU and QR Factorization

Let S be a real N by N matrix. Two important methods for solving the system $Sx = b$, the LU factorization and the QR factorization, involve factoring the matrix S and thereby reducing the problem to finding the solutions of simpler systems.

In the LU factorization, we seek a lower triangular matrix L and an upper triangular matrix U so that $S = LU$. We then solve $Sx = b$ by solving $Lz = b$ and $Ux = z$.

In the QR factorization, we seek an orthogonal matrix Q, that is, $Q^T = Q^{-1}$, and an upper triangular matrix R so that $S = QR$. Then we solve $Sx = b$ by solving the upper triangular system $Rx = Q^T b$.

23.2 The LU Factorization

The matrix

$$S = \begin{bmatrix} 2 & 1 & 1 \\ 4 & 1 & 0 \\ -2 & 2 & 1 \end{bmatrix}$$

can be reduced to the upper triangular matrix

$$U = \begin{bmatrix} 2 & 1 & 1 \\ 0 & -1 & -2 \\ 0 & 0 & -4 \end{bmatrix}$$

through three elementary row operations: first, add -2 times the first row to the second row; second, add the first row to the third row; finally, add three times the new second row to the third row. Each of these row operations can be viewed as the result of multiplying on the left by the matrix obtained by applying the same row operation to the identity matrix. For example, adding -2 times the first row to the second row can be achieved by multiplying A on the left by the matrix

$$L_1 = \begin{bmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix};$$

note that the inverse of L_1 is

$$L_1^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

We can write

$$L_3 L_2 L_1 S = U,$$

where L_1, L_2, and L_3 are the matrix representatives of the three elementary row operations. Therefore, we have

$$S = L_1^{-1} L_2^{-1} L_3^{-1} U = LU.$$

This is the *LU factorization* of S. As we just saw, the LU factorization can be obtained along with the Gauss elimination.

23.2.1 A Shortcut

There is a shortcut we can take in calculating the LU factorization. We begin with the identity matrix I, and then, as we perform a row operation,

for example, adding -2 times the first row to the second row, we put the number 2, the multiplier just used, but with a sign change, in the second row, first column, the position of the entry of S that was just converted to zero. Continuing in this fashion, we build up the matrix L as

$$L = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ -1 & -3 & 1 \end{bmatrix},$$

so that

$$S = \begin{bmatrix} 2 & 1 & 1 \\ 4 & 1 & 0 \\ -2 & 2 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ -1 & -3 & 1 \end{bmatrix} \begin{bmatrix} 2 & 1 & 1 \\ 0 & -1 & -2 \\ 0 & 0 & -4 \end{bmatrix}.$$

The entries of the main diagonal of L will be all ones. If we want the same to be true of U, we can rescale the rows of U and obtain the factorization $S = LDU$, where D is a diagonal matrix.

23.2.2 A Warning!

We have to be careful when we use the shortcut, as we illustrate now. For the purpose of this discussion let's use the terminology $R_i + aR_j$ to mean the row operation that adds a times the jth row to the ith row, and aR_i to mean the operation that multiplies the ith row by a. Now we transform S to an upper triangular matrix U using the row operations

- 1. $\frac{1}{2}R_1$;
- 2. $R_2 + (-4)R_1$;
- 3. $R_3 + 2R_1$;
- 4. $R_3 + 3R_2$;
- 5. $(-1)R_2$; and finally,
- 6. $(\frac{-1}{4})R_3$.

We end up with

$$U = \begin{bmatrix} 1 & 1/2 & 1/2 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{bmatrix}.$$

If we use the shortcut to form the lower triangular matrix L, we find that

$$L = \begin{bmatrix} 2 & 0 & 0 \\ 4 & -1 & 0 \\ -2 & -3 & -4 \end{bmatrix}.$$

Let's go through how we formed L from the row operations listed above. We get $L_{11} = 2$ from the first row operation, $L_{21} = 4$ from the second, $L_{31} = -2$ from the third, $L_{32} = -3$ from the fourth, $L_{22} = -1$ from the fifth, and $L_{33} = \frac{-1}{4}$ from the sixth. But, if we multiple LU we do not get back S! The problem is that we performed the fourth operation, adding to the third row three times the second row, before the $(2, 2)$ entry was rescaled to one. Suppose, instead, we do the row operations in this order:

- 1. $\frac{1}{2}R_1$;
- 2. $R_2 + (-4)R_1$;
- 3. $R_3 + 2R_1$;
- 4. $(-1)R_2$;
- 5. $R_3 - 3R_2$; and finally,
- 6. $(\frac{-1}{4})R_3$.

Then the entry L_{32} becomes 3, instead of -3, and now $LU = S$. The message is that if we want to use the shortcut and we plan to rescale the diagonal entries of U to be one, we should rescale a given row prior to adding any multiple of that row to another row; otherwise, we can get the wrong L. The problem is that certain elementary matrices associated with row operations do not commute.

We just saw that

$$L = L_1^{-1}L_2^{-1}L_3^{-1}.$$

However, when we form the matrix L simultaneously with performing the row operations, we are, in effect, calculating

$$L_3^{-1}L_2^{-1}L_1^{-1}.$$

Most of the time the order doesn't matter, and we get the correct L anyway. But this is not always the case. For example, if we perform the operation $\frac{1}{2}R_1$, followed by $R_2 + (-4)R_1$, this is not the same as doing $R_2 + (-4)R_1$, followed by $\frac{1}{2}R_1$.

With the matrix L_1 representing the operation $\frac{1}{2}R_1$ and the matrix L_2 representing the operation $R_2 + (-4)R_1$, we find that storing a 2 in the $(1, 1)$ position, and then a $+4$ in the $(1, 2)$ position as we build L is not equivalent to multiplying the identity matrix by $L_2^{-1}L_1^{-1}$ but rather multiplying the identity matrix by

$$(L_1^{-1}L_2^{-1}L_1)L_1^{-1} = L_1^{-1}L_2^{-1},$$

which is the correct order.

To illustrate this point, consider the matrix S given by

$$S = \begin{bmatrix} 2 & 1 & 1 \\ 4 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

In the first instance, we perform the row operations $R_2 + (-2)R_1$, followed by $\frac{1}{2}R_1$ to get

$$U = \begin{bmatrix} 1 & 0.5 & 0.5 \\ 0 & -1 & -2 \\ 0 & 0 & 1 \end{bmatrix}.$$

Using the shortcut, the matrix L becomes

$$L = \begin{bmatrix} 2 & 0 & 0 \\ 2 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

but we do not get $S = LU$. We do have $U = L_2L_1S$, where

$$L_1 = \begin{bmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

and

$$L_2 = \begin{bmatrix} 0.5 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

so that $S = L_1^{-1}L_2^{-1}U$ and the correct L is

$$L = L_1^{-1}L_2^{-1} = \begin{bmatrix} 2 & 0 & 0 \\ 4 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

But when we use the shortcut to generate L, we effectively multiply the identity matrix first by L_1^{-1} and then by L_2^{-1}, giving the matrix $L_2^{-1}L_1^{-1}$ as our candidate for L. But $L_1^{-1}L_2^{-1}$ and $L_2^{-1}L_1^{-1}$ are not the same. But why does reversing the order of the row operations work?

When we perform $\frac{1}{2}R_1$ first, and then $R_2 + (-4)R_1$ to get U, we are multiplying S first by L_2 and then by the matrix

$$E = \begin{bmatrix} 1 & 0 & 0 \\ -4 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

The correct L is then $L = L_2^{-1}E^{-1}$.

When we use the shortcut, we are first multiplying the identity by the matrix L_2^{-1} and then by a second matrix that we shall call J; the correct L must then be $L = JL_2^{-1}$. The matrix J is not E^{-1}, but

$$J = L_2^{-1}E^{-1}L_2,$$

so that

$$L = J - +L_2^{-1} = L_2^{-1}E^{-1}L_2L_2^{-1} = L_2^{-1}E^{-1},$$

which is correct.

23.2.3 Using the LU decomposition

Suppose that we have to solve the system of linear equations $Sx = b$. Once we have the LU factorization, it is a simple matter to find x: first, we solve the system $Lz = b$, and then solve $Ux = z$. Because both L and U are triangular, solving these systems is a simple matter. Obtaining the LU factorization is often better than finding S^{-1}; when S is banded, that is, has non-zero values only for the main diagonal and a few diagonals on either side, the L and U retain that banded property, while S^{-1} does not.

If H is real and symmetric, and if $H = LDU$, then $U = L^T$, so we have $H = LDL^T$. If, in addition, the non-zero entries of D are positive, then we can write

$$H = (L\sqrt{D})(L\sqrt{D})^T,$$

which is the Cholesky Decomposition of H.

Ex. 23.1 *Prove that if L is invertible and lower triangular , then so is L^{-1}.*

Ex. 23.2 *Show that the symmetric matrix*

$$H = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

cannot be written as $H = LDL^T$.

Ex. 23.3 *Show that the symmetric matrix*

$$H = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

cannot be written as $H = LU$, where L is lower triangular, U is upper triangular, and both are invertible.

23.2.4 The Non-Square Case

If A is an M by N matrix, the same elimination procedure leads to a factoring $PA = LU$, where now the matrix L is square and lower-triangular and the matrix U is in *upper echelon form*, meaning that

- 1. the non-zero rows of U come at the top of U and the first non-zero entries are called the *pivots*;
- 2. below each pivot is a column of zeros;
- 3. each pivot lies to the right of the pivot in the row above it.

23.2.5 The LU Factorization in Linear Programming

Each step of the simplex algorithm involves solving systems of equations of the form $Bx = b$ and $B^T z = c$. As we proceed from one step to the next, the matrix B is updated by having one of its columns changed. This can be performed by multiplying B on the right by a matrix F that is the identity matrix, except for one column. The matrix $E = F^{-1}$ is then also the identity matrix, except for one column, so the updated inverse is

$$(B^{\text{new}})^{-1} = EB^{-1}.$$

As the calculations proceed, the next inverse can be represented in product form as

$$(B^{\text{new}})^{-1} = E_k E_{k-1} \cdots E_1 (B_0)^{-1},$$

where B_0 is the original choice for the matrix B. This product approach suggests a role for LU factorization, in which the individual factors L and U are updated in a stable manner as the iteration proceeds [266].

Ex. 23.4

- *a. Show that the matrix $B = A + x\delta_n^T$ differs from A only in the nth column, where x is an arbitrary column vector and δ_n is the nth column of the identity matrix.*
- *b. Let F be a matrix that is the identity matrix, except for one column. Show that the matrix $E = F^{-1}$, when it exists, is then also the identity matrix, except for one column, and compute E explicitly, in terms of the entries of F.*

Hint: use the identity in Equation 3.5.

23.3 When is $S = LU$?

Note that it may not be possible to obtain $S = LU$ without first permuting the rows of S; in such cases we obtain $PS = LU$, where P is obtained from the identity matrix by permuting rows.

We know from Exercise 23.3 that the invertible symmetric matrix

$$H = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

cannot be written as $H = LU$, where both L and U are invertible. In [269] Mark Yin gave a necessary and sufficient condition for a square matrix S to have the form $S = LU$, where both L and U are invertible.

Definition 23.1 *An n by n real matrix S is called a* T-matrix *if, for every partition*

$$S = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix}$$

such that S_{11} is square, S_{11} is invertible.

Yin's theorem is the following:

Theorem 23.1 *An n by n matrix S has the form $S = LU$, where L is lower triangular, U is upper triangular, and both are invertible, if and only if S is a T-matrix.*

Proof: Suppose that $S = LU$ as in the statement of the theorem. Let S be partitioned arbitrarily, as

$$S = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix},$$

where S_{11} is square. Let

$$P = L^{-1} = \begin{bmatrix} P_{11} & 0 \\ P_{21} & P_{22} \end{bmatrix}$$

be an invertible lower triangular matrix, partitioned to be compatible with the partitioning of S, such that the matrix

$$PS = \begin{bmatrix} P_{11}S_{11} & P_{11}S_{12} \\ 0 & * \end{bmatrix}$$

is invertible and upper triangular. Since $P_{11}S_{11}$ must then be invertible, so is S_{11}. Since the partition is arbitrary, S is a T-matrix.

Now suppose that S is a T-matrix. We show that $S = LU$ as above. First of all, notice that, if P is invertible and lower triangular, then PS is also a T-matrix, since the upper left corner square sub-matrix of PS is $P_{11}S_{11}$.

The proof uses induction on the size n. The case of $n = 1$ is trivial, so assume that $n > 1$ and that the theorem holds for square matrices of size $n-1$ by $n-1$. Let

$$P_1 = \begin{bmatrix} 1 & 0^T \\ b & I \end{bmatrix},$$

where I is the identity matrix, s_{mn} are the entries of the matrix S, and

$$b^T = -\frac{1}{s_{11}}(s_{21}, ..., s_{n1}).$$

Then

$$P_1 S = \begin{bmatrix} s_{11} & 0^T \\ S_{12} & A_{22} \end{bmatrix},$$

where A_{22} is square and has size $n-1$. Since P_1S is a T-matrix, so is A_{22}. By the induction hypothesis, there is an invertible lower triangular matrix P_2 such that P_2A_{22} is invertible and upper triangular. It follows that RP_1S is invertible and upper triangular, where

$$R = \begin{bmatrix} 1 & 0^T \\ 0 & P_2 \end{bmatrix}.$$

Since RP_1 is invertible and lower triangular, the proof is completed. ∎

23.4 Householder Matrices

A real *Householder matrix* has the form

$$H = I - 2ww^T,$$

where w is a column vector in $\mathbb{R}^N$ with $\|w\|_2 = 1$.

Lemma 23.1 *For any Householder matrix we have $H^T = H$ and $H^{-1} = H$.*

Ex. 23.5 *Prove Lemma 23.1.*

Proposition 23.1 *Let x and y be any members of $\mathbb{R}^N$ with $\|x\|_2 = \|y\|_2$, and let $w = \frac{1}{\|x-y\|_2}(x-y)$. Then $Hx = y$.*

Ex. 23.6 *Prove Proposition 23.1.*

We can use Householder matrices to turn certain non-zero entries of a vector to zero.

Given any vector x in $\mathbb{R}^N$, let $y_n = x_n$, for $n = 1, ..., k-1$, $y_n = 0$, for $n = k+1, ..., N$, and

$$|y_k| = \sqrt{x_k^2 + x_{k+1}^2 + ... + x_N^2},$$

where the sign of y_k is chosen to be opposite that of x_k. Then $\|x\|_2 = \|y\|_2$, the first $k-1$ entries of x and y agree, and the final $N-k$ entries of y are zero. If we then build the Householder matrix H using these x and y to create w, we find that $Hx = y$, so that the final $N-k$ entries are zero.

23.5 The QR Factorization

Given an invertible N by N real matrix S, we can multiply S on the left by a succession of Householder matrices $H_1, H_2, ..., H_{k-1}$ so that

$$H_{k-1} \cdots H_1 S = R$$

is upper triangular. Since $H_n^T = H_n = H_n^{-1}$, it follows that

$$Q^T = H_{k-1} \cdots H_1$$

is orthogonal, and that $S = QR$. This is the QR factorization of S. Once we have $S = QR$, we can solve $Sx = b$ easily, by solving $Rx = Q^T b$.

23.5.1 The Non-Square Case

Using the same approach, any real rectangular matrix A with linearly independent columns can be factored as $A = QR$, where R is square, upper triangular, and invertible, and the columns of Q are orthonormal, so that $Q^T Q = I$.

23.5.2 The QR Factorization and Least Squares

The least-squares solution of $Ax = b$ is the solution of $A^T Ax = A^T b$. Once we have $A = QR$, we have $A^T A = R^T Q^T QR = R^T R$, so we find the least squares solution easily, by solving $R^T z = A^T b$, and then $Rx = z$. Note that $A^T A = R^T R$ is the Cholesky decomposition of $A^T A$.

23.5.3 Upper Hessenberg Matrices

The time required to calculate the QR factorization of a general N by N matrix is proportional to N^3; the time is proportional to N^2 if the matrix has the *upper Hessenberg* form.

We say that a real N by N matrix has upper Hessenberg form if its non-zero entries occur on or above the main diagonal (as with an upper triangular matrix), or on the first sub-diagonal below the main diagonal. Note that any real N by N matrix S can be converted to upper Hessenberg form by multiplying on the left by a succession of Householder matrices; we can find Householder matrices H_1, H_2, ..., H_{k-2} so that

$$H_{k-2} \cdots H_1 A = B,$$

with B in upper Hessenberg form. The matrix

$$C = BH_1 \cdots H_{k-2} = H_{k-2} \cdots H_1 A H_1 \cdots H_{k-2}$$

is also in upper Hessenberg form. Since $C = P^{-1}AP$ for an invertible matrix P, the matrix C is similar to A, and so has the same eigenvalues. This will be helpful later.

23.5.4 The QR Method for Finding Eigenvalues

The QR factorization can be used to calculate the eigenvalues of a real N by N matrix S. The method proceeds as follows: begin with $S = S_0 = Q_0 R_0$, then define $S_1 = R_0 Q_0$. Next, perform the QR factorization on S_1 to get $S_1 = Q_1 R_1$, and define $S_2 = R_1 Q_1$, and so on. If S has only real eigenvalues, this procedure usually converges to an upper triangular matrix, whose eigenvalues are displayed along its main diagonal. Since $S_k = Q_k R_k$ and Q_k is orthogonal, we have $R_k = (Q_k)^T S_k$, so that

$$S_{k+1} = R_k Q_k = (Q_k)^T S_k Q_k = (Q_k)^{-1} S_k Q_k.$$

Therefore, each S_k is similar to S and so they have the same eigenvalues.

Chapter 24

Appendix: More ART and MART

24.1 Chapter Summary

Although the ART and the MART were developed to compute tomographic images, they can be viewed more generally as iterative methods for solving systems of linear equations.

24.2 The ART in the General Case

Let A be a complex matrix with I rows and J columns, and let b be a member of $\mathbb{C}^I$. We want to solve the system $Ax = b$. For each index value i, let H_i be the hyperplane of J-dimensional vectors given by

$$H_i = \{x | (Ax)_i = b_i\}, \tag{24.1}$$

and P_i the orthogonal projection operator onto H_i. Let x^0 be arbitrary and, for each nonnegative integer k, let $i(k) = k(\mathrm{mod}\, I) + 1$. The iterative step of the ART is

$$x^{k+1} = P_{i(k)}x^k. \tag{24.2}$$

Because the ART uses only a single equation at each step, it has been called a *row-action* method .

24.2.1 Calculating the ART

Given any vector z the vector in H_i closest to z, in the sense of the Euclidean distance, has the entries

$$x_j = z_j + \overline{A_{ij}}(b_i - (Az)_i)/\sum_{m=1}^{J} |A_{im}|^2. \tag{24.3}$$

To simplify our calculations, we shall assume, throughout this chapter, that the rows of A have been rescaled to have Euclidean length one; that is

$$\sum_{j=1}^{J} |A_{ij}|^2 = 1, \tag{24.4}$$

for each $i = 1, ..., I$, and that the entries of b have been rescaled accordingly, to preserve the equations $Ax = b$. The ART is then the following: begin with an arbitrary vector x^0; for each nonnegative integer k, having found x^k, the next iterate x^{k+1} has entries

$$x_j^{k+1} = x_j^k + \overline{A_{ij}}(b_i - (Ax^k)_i). \tag{24.5}$$

As we shall show, when the system $Ax = b$ has exact solutions the ART converges to the solution closest to x^0, in the 2-norm. How fast the algorithm converges will depend on the ordering of the equations and on whether or not we use relaxation. In selecting the equation ordering, the important thing is to avoid particularly bad orderings, in which the hyperplanes H_i and H_{i+1} are nearly parallel.

24.2.2 Full-cycle ART

We again consider the *full-cycle* ART, with iterative step $z^{m+1} = Tz^m$, for

$$T = P_I P_{I-1} \cdots P_2 P_1. \tag{24.6}$$

When the system $Ax = b$ has solutions, the fixed points of T are solutions. When there are no solutions of $Ax = b$, the operator T will still have fixed points, but they will no longer be exact solutions.

24.2.3 Relaxed ART

The ART employs orthogonal projections onto the individual hyperplanes. If we permit the next iterate to fall short of the hyperplane, or somewhat beyond it, we get a relaxed version of ART.The relaxed ART algorithm is as follows:

Algorithm 24.1 (Relaxed ART) *With $\omega \in (0, 2)$, x^0 arbitrary, and $i = k(\mathrm{mod}\, I) + 1$, let*

$$x_j^{k+1} = x_j^k + \omega \overline{A_{ij}}(b_i - (Ax^k)_i)). \tag{24.7}$$

The relaxed ART converges to the solution closest to x^0, in the consistent case. In the inconsistent case, it does not converge, but subsequences associated with the same i converge to distinct vectors, forming a limit cycle.

24.2.4 Constrained ART

Let C be a closed, nonempty convex subset of $\mathbb{C}^J$ and $P_C x$ the orthogonal projection of x onto C. If there are solutions of $Ax = b$ that lie within C, we can find them using the constrained ART algorithm:

Algorithm 24.2 (Constrained ART) *With x^0 arbitrary and $i = k(\mathrm{mod}\, I) + 1$, let*

$$z_j^{k+1} = x_j^k + \overline{A_{ij}}(b_i - (Ax^k)_i), \tag{24.8}$$

and $x^{k+1} = P_C z^{k+1}$.

For example, if A and b are real and we seek a nonnegative solution to $Ax = b$, we can use

Algorithm 24.3 (Non-negative ART) *With $i = k(\mathrm{mod}\, I) + 1$, and x^0 arbitrary, let*

$$x_j^{k+1} = (x_j^k + A_{ij}(b_i - (Ax^k)_i))_+, \tag{24.9}$$

where, for any real number a, $a_+ = \max\{a, 0\}$.

The constrained ART converges to a solution of $Ax = b$ within C, whenever such solutions exist.

Noise in the data vector b can manifest itself in a variety of ways. Suppose that the system $Ax = b$ ought to have non-negative solutions, but because the entries of b are noisy measured data, it does not. Theorem 24.1 tells us that when $J > I$, but $Ax = b$ has no non-negative solutions, the non-negatively constrained least-squares solution can have at most $I - 1$

non-zero entries, regardless of how large J is. This phenomenon also occurs with several other approximate methods, such as those that minimize the cross-entropy distance. This gives us a sense of what can happen when we impose positivity on the calculated least-squares solution, that is, when we minimize $||Ax - b||_2$ over all non-negative vectors x.

Definition 24.1 *The matrix A has the* full-rank property *if A and every matrix Q obtained from A by deleting columns have full rank.*

Theorem 24.1 *Let A have the full-rank property. Suppose there is no non-negative solution to the system of equations $Ax = b$. Then there is a subset S of the set $\{j = 1, 2, ..., J\}$, with cardinality at most $I - 1$, such that, if $\hat{x}$ is any minimizer of $||Ax - b||_2$ subject to $x \geq 0$, then $\hat{x}_j = 0$ for j not in S. Therefore, $\hat{x}$ is unique.*

For a proof, see the chapter on optimization.

24.2.5 When $Ax = b$ Has Solutions

For the consistent case, in which the system $Ax = b$ has exact solutions, we have the following result.

Theorem 24.2 *Let $A\hat{x} = b$ and let x^0 be arbitrary. Let $\{x^k\}$ be generated by Equation (24.5). Then the sequence $\{||\hat{x} - x^k||_2\}$ is decreasing and $\{x^k\}$ converges to the solution of $Ax = b$ closest to x^0.*

The proof of the following lemma follows from the definition of the ART iteration, with a little algebraic manipulation.

Lemma 24.1 *Let x^0 and y^0 be arbitrary and $\{x^k\}$ and $\{y^k\}$ be the sequences generated by applying the ART algorithm, beginning with x^0 and y^0, respectively; that is, $y^{k+1} = P_{i(k)}y^k$. Then*

$$||x^0 - y^0||_2^2 - ||x^I - y^I||_2^2 = \sum_{i=1}^{I} |(Ax^{i-1})_i - (Ay^{i-1})_i|^2. \qquad (24.10)$$

Ex. 24.1 *Prove Lemma 24.1.*

Proof of Theorem 24.2: Let $A\hat{x} = b$. Let $v_i^r = (Ax^{rI+i-1})_i$ and $v^r = (v_1^r, ..., v_I^r)^T$, for $r = 0, 1, ...$ It follows from Equation (24.10) that the sequence $\{||\hat{x} - x^{rI}||_2\}$ is decreasing and the sequence $\{v^r - b\} \to 0$. So $\{x^{rI}\}$ is bounded; let $x^{*,0}$ be a cluster point. Then, for $i = 1, 2, ..., I$, let $x^{*,i}$ be the successor of $x^{*,i-1}$ using the ART algorithm. It follows that $(Ax^{*,i-1})_i = b_i$ for each i, from which we conclude that $x^{*,0} = x^{*,i}$ for all

i and that $Ax^{*,0} = b$. Using $x^{*,0}$ in place of the arbitrary solution $\hat{x}$, we have that the sequence $\{||x^{*,0} - x^k||_2\}$ is decreasing. But a subsequence converges to zero, so $\{x^k\}$ converges to $x^{*,0}$. By Equation (24.10), the difference $||\hat{x} - x^k||_2^2 - ||\hat{x} - x^{k+1}||_2^2$ is independent of which solution $\hat{x}$ we pick; consequently, so is $||\hat{x} - x^0||_2^2 - ||\hat{x} - x^{*,0}||_2^2$. It follows that $x^{*,0}$ is the solution closest to x^0. This completes the proof. ∎

24.2.6 When $Ax = b$ Has No Solutions

When there are no exact solutions, the ART does not converge to a single vector, but, for each fixed i, the subsequence $\{x^{nI+i}, n = 0, 1, ...\}$ converges to a vector z^i and the collection $\{z^i \, | i = 1, ..., I\}$ is called the *limit cycle*. This was shown by Tanabe [247] and also follows from the results of De Pierro and Iusem [111]. Proofs of subsequential convergence are given in [65, 66]. The ART limit cycle will vary with the ordering of the equations, and contains more than one vector unless an exact solution exists.

Open Question: If there is a unique geometric least-squares solution, where is it, in relation to the vectors of the limit cycle? Can it be calculated easily, from the vectors of the limit cycle?

There is a partial answer to the second question. In [55] (see also [65]) it was shown that if the system $Ax = b$ has no exact solution, and if $I = J+1$, then the vectors of the limit cycle lie on a sphere in J-dimensional space having the least-squares solution at its center. This is not true more generally, however.

Open Question: In both the consistent and inconsistent cases, the sequence $\{x^k\}$ of ART iterates is bounded, as Tanabe [247], and De Pierro and Iusem [111] have shown. The proof is easy in the consistent case. Is there an easy proof for the inconsistent case?

24.3 Regularized ART

If the entries of b are noisy but the system $Ax = b$ remains consistent (which can easily happen in the under-determined case, with $J > I$), the ART begun at $x^0 = 0$ converges to the solution having minimum Euclidean norm, but this norm can be quite large. The resulting solution is probably useless.

We know from a previous exercise that the system $AA^\dagger z = b$ has a solution if and only if the system $Ax = b$ has solutions.

Ex. 24.2 *Show that the matrix $AA^\dagger+\epsilon I$ is always invertible, for any $\epsilon > 0$. Then show that*

$$(AA^\dagger + \epsilon I)^{-1}A = A(A^\dagger A + \epsilon I)^{-1}.$$

Instead of solving $Ax = b$, we *regularize* by minimizing, for example, the function

$$F_\epsilon(x) = ||Ax - b||_2^2 + \epsilon^2||x||_2^2. \tag{24.11}$$

The solution to this problem is the vector

$$\hat{x}_\epsilon = (A^\dagger A + \epsilon^2 I)^{-1}A^\dagger b, \tag{24.12}$$

which always exists, even when the system $Ax = b$ has no solutions.

However, we do not want to calculate $A^\dagger A + \epsilon^2 I$ when the matrix A is large. Fortunately, there are ways to find $\hat{x}_\epsilon$, using only the matrix A and the ART algorithm.

We discuss two methods for using ART to obtain regularized solutions of $Ax = b$. The first one is presented in [65], while the second one is due to Eggermont, Herman, and Lent [124].

Both methods rely on the fact that when the ART is applied to a consistent system $Ax = b$ it converges to the solution of that system closest to where we began the iteration. We know from Theorem 3.2 that the solution of $Ax = b$ closest to the origin has the form $x = A^\dagger z$, so that $b = AA^\dagger z$. Assuming $AA^\dagger$ is invertible, we have $z = (AA^\dagger)^{-1}b$ and and

$$x = A^\dagger(AA^\dagger)^{-1}b.$$

If we want to find the solution closest to a given vector p, we write $t = x-p$, so that $At = Ax - Ap = b - Ap$ and then find the solution of $At = b - Ap$ closest to the origin. Then

$$t = A^\dagger(AA^\dagger)^{-1}(b - Ap),$$

and

$$x = t + p = A^\dagger(AA^\dagger)^{-1}(b - Ap) + p.$$

In our first method we use ART to solve the system of equations given in matrix form by

$$\begin{bmatrix} A^\dagger & \epsilon I \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = 0. \tag{24.13}$$

We begin with $u^0 = b$ and $v^0 = 0$. Then, the lower component of the limit vector is $v^\infty = -\epsilon\hat{x}_\epsilon$.

The method of Eggermont *et al.* is similar. In their method we use ART to solve the system of equations given in matrix form by

$$\begin{bmatrix} A & \epsilon I \end{bmatrix} \begin{bmatrix} x \\ v \end{bmatrix} = b. \tag{24.14}$$

We begin at $x^0 = 0$ and $v^0 = 0$. Then, the limit vector has for its upper component $x^\infty = \hat{x}_\epsilon$ as before, and that $\epsilon v^\infty = b - A\hat{x}_\epsilon$.

Ex. 24.3 *Prove that the two iterative methods for regularized ART perform as indicated.*

24.4 Avoiding the Limit Cycle

Generally, the greater the minimum value of $||Ax - b||_2^2$ the more the vectors of the LC are distinct from one another. There are several ways to avoid the LC in ART and to obtain a least-squares solution. One way is the *double ART* (DART) [59]:

24.4.1 Double ART (DART)

We know that any b can be written as $b = A\hat{x} + \hat{w}$, where $A^T\hat{w} = 0$ and $\hat{x}$ is a minimizer of $||Ax - b||_2^2$. The vector $\hat{w}$ is the orthogonal projection of b onto the null space of the matrix transformation $A^\dagger$. Therefore, in Step 1 of DART we apply the ART algorithm to the consistent system of linear equations $A^\dagger w = 0$, beginning with $w^0 = b$. The limit is $w^\infty = \hat{w}$, the member of the null space of $A^\dagger$ closest to b. In Step 2, apply ART to the consistent system of linear equations $Ax = b - w^\infty = A\hat{x}$. The limit is then the minimizer of $||Ax - b||_2$ closest to x^0. Notice that we could also obtain the least-squares solution by applying ART to the system $A^\dagger y = A^\dagger b$, starting with $y^0 = 0$, to obtain the minimum-norm solution, which is $y = A\hat{x}$, and then applying ART to the system $Ax = y$.

24.4.2 Strongly Under-relaxed ART

Another method for avoiding the LC is *strong under-relaxation*, due to Censor, Eggermont and Gordon [80]. Let $t > 0$. Replace the iterative step in ART with

$$x_j^{k+1} = x_j^k + t\overline{A_{ij}}(b_i - (Ax^k)_i). \tag{24.15}$$

In [80] it is shown that, as $t \to 0$, the vectors of the LC approach the geometric least squares solution closest to x^0; a short proof is in [55]. Bertsekas [20] uses strong under-relaxation to obtain convergence of more general incremental methods.

24.5 The MART

The *multiplicative* ART (MART) [151] is an iterative algorithm closely related to the ART. It also was devised to obtain tomographic images, but, like ART, applies more generally; MART applies to systems of linear equations $Ax = b$ for which the b_i are positive, the A_{ij} are nonnegative, and the solution x we seek is to have nonnegative entries. It is not so easy to see the relation between ART and MART if we look at the most general formulation of MART. For that reason, we began with a simpler case, transmission tomographic imaging, in which the relation is most clearly visible.

24.5.1 The MART in the General Case

The MART, which can be applied only to nonnegative systems, is a sequential, or row-action, method that uses one equation only at each step of the iteration.

Algorithm 24.4 (MART) *Let x^0 be any positive vector, and $i = k(\mathrm{mod}\, I) + 1$. Having found x^k for positive integer k, define x^{k+1} by*

$$x_j^{k+1} = x_j^k \Big(\frac{b_i}{(Ax^k)_i} \Big)^{m_i^{-1} A_{ij}}, \tag{24.16}$$

where $m_i = \max \{A_{ij} \,|j = 1, 2, ..., J\}$.

Some treatments of MART leave out the m_i, but require only that the entries of A have been rescaled so that $A_{ij} \leq 1$ for all i and j. The m_i is important, however, in accelerating the convergence of MART. There is another way to do the rescaling for MART, which we discuss in the appendix on Geometric Programming and the MART.

The MART can be accelerated by relaxation, as well.

Algorithm 24.5 (Relaxed MART) *Let x^0 be any positive vector, and $i = k(\mathrm{mod}\, I) + 1$. Having found x^k for positive integer k, define x^{k+1} by*

$$x_j^{k+1} = x_j^k \Big(\frac{b_i}{(Ax^k)_i} \Big)^{\gamma_i m_i^{-1} A_{ij}}, \tag{24.17}$$

where γ_i is in the interval $(0, 1)$.

As with ART, finding the best relaxation parameters is a bit of an art.

24.5.2 Cross-Entropy

For $a > 0$ and $b > 0$, let the cross-entropy or Kullback-Leibler distance from a to b be

$$KL(a, b) = a \log \frac{a}{b} + b - a, \tag{24.18}$$

with $KL(a, 0) = +\infty$, and $KL(0, b) = b$. Extend to nonnegative vectors coordinate-wise, so that

$$KL(x, z) = \sum_{j=1}^{J} KL(x_j, z_j). \tag{24.19}$$

Unlike the Euclidean distance, the KL distance is not symmetric; $KL(Ax, b)$ and $KL(b, Ax)$ are distinct, and we can obtain different approximate solutions of $Ax = b$ by minimizing these two distances with respect to non-negative x.

24.5.3 Convergence of MART

In the consistent case, by which we mean that $Ax = b$ has nonnegative solutions, we have the following convergence theorem for MART.

Theorem 24.3 *In the consistent case, the MART converges to the unique nonnegative solution of $b = Ax$ for which the distance $\sum_{j=1}^{J} KL(x_j, x_j^0)$ is minimized.*

If the starting vector x^0 is the vector whose entries are all one, then the MART converges to the solution that maximizes the Shannon entropy,

$$SE(x) = \sum_{j=1}^{J} x_j \log x_j - x_j. \tag{24.20}$$

As with ART, the speed of convergence is greatly affected by the ordering of the equations, converging most slowly when consecutive equations correspond to nearly parallel hyperplanes.

Open Question: When there are no nonnegative solutions, MART does not converge to a single vector, but, like ART, is always observed to produce a limit cycle of vectors. Unlike ART, there is no proof of the existence of a limit cycle for MART.

Chapter 25

Appendix: Constrained Iteration Methods

25.1 Chapter Summary

The ART and its simultaneous and block-iterative versions are designed to solve general systems of linear equations $Ax = b$. The SMART, EMML and RBI methods require that the entries of A be nonnegative, those of b positive and produce nonnegative x. In this chapter we present variations of the SMART and EMML that impose the constraints $u_j \leq x_j \leq v_j$, where the u_j and v_j are selected lower and upper bounds on the individual entries x_j. These algorithms were used in [208] as a method for including in transmission tomographic reconstruction spatially varying upper and lower bounds on the x-ray attenuation.

25.2 Modifying the KL distance

The SMART, EMML and RBI methods are based on the Kullback-Leibler distance between nonnegative vectors. To impose more general constraints on the entries of x we derive algorithms based on shifted KL distances, also called Fermi-Dirac generalized entropies.

For a fixed real vector u, the shifted KL distance $KL(x - u, z - u)$ is defined for vectors x and z having $x_j \geq u_j$ and $z_j \geq u_j$. Similarly, the shifted distance $KL(v - x, v - z)$ applies only to those vectors x and z for

which $x_j \leq v_j$ and $z_j \leq v_j$. For $u_j \leq v_j$, the combined distance

$$KL(x-u, z-u) + KL(v-x, v-z)$$

is restricted to those x and z whose entries x_j and z_j lie in the interval $[u_j, v_j]$. Our objective is to mimic the derivation of the SMART, EMML and RBI methods, replacing KL distances with shifted KL distances, to obtain algorithms that enforce the constraints $u_j \leq x_j \leq v_j$, for each j. The algorithms that result are the ABMART and ABEMML block-iterative methods. These algorithms were originally presented in [57], in which the vectors u and v were called a and b, hence the names of the algorithms. Throughout this chapter we shall assume that the entries of the matrix A are nonnegative. We shall denote by B_n, $n = 1, ..., N$ a partition of the index set $\{i = 1, ..., I\}$ into blocks. For $k = 0, 1, ...$ let $n(k) = k(\text{mod}\, N) + 1$.

The projected Landweber algorithm can also be used to impose the restrictions $u_j \leq x_j \leq v_j$; however, the projection step in that algorithm is implemented by clipping, or setting equal to u_j or v_j values of x_j that would otherwise fall outside the desired range. The result is that the values u_j and v_j can occur more frequently than may be desired. One advantage of the AB methods is that the values u_j and v_j represent barriers that can only be reached in the limit and are never taken on at any step of the iteration.

25.3 The ABMART Algorithm

We assume that $(Au)_i \leq b_i \leq (Av)_i$ and seek a solution of $Ax = b$ with $u_j \leq x_j \leq v_j$, for each j. The algorithm begins with an initial vector x^0 satisfying $u_j \leq x_j^0 \leq v_j$, for each j. Having calculated x^k, we take

$$x_j^{k+1} = \alpha_j^k v_j + (1 - \alpha_j^k) u_j, \tag{25.1}$$

with $n = n(k)$,

$$\alpha_j^k = \frac{c_j^k \prod^n (d_i^k)^{A_{ij}}}{1 + c_j^k \prod^n (d_i^k)^{A_{ij}}}, \tag{25.2}$$

$$c_j^k = \frac{(x_j^k - u_j)}{(v_j - x_j^k)}, \tag{25.3}$$

and

$$d_j^k = \frac{(b_i - (Au)_i)((Av)_i - (Ax^k)_i)}{((Av)_i - b_i)((Ax^k)_i - (Au)_i)}, \tag{25.4}$$

where $\prod^n$ denotes the product over those indices i in $B_{n(k)}$. Notice that, at each step of the iteration, x_j^k is a convex combination of the endpoints u_j and v_j, so that x_j^k lies in the interval $[u_j, v_j]$.

We have the following theorem concerning the convergence of the AB-MART algorithm:

Theorem 25.1 *If there is a solution of the system $Ax = b$ that satisfies the constraints $u_j \leq x_j \leq v_j$ for each j, then, for any N and any choice of the blocks B_n, the ABMART sequence converges to that constrained solution of $Ax = b$ for which the Fermi-Dirac generalized entropic distance from x to x^0,*

$$KL(x - u, x^0 - u) + KL(v - x, v - x^0),$$

is minimized. If there is no constrained solution of $Ax = b$, then, for $N = 1$, the ABMART sequence converges to the minimizer of

$$KL(Ax - Au, b - Au) + KL(Av - Ax, Av - b)$$

for which

$$KL(x - u, x^0 - u) + KL(v - x, v - x^0)$$

is minimized.

The proof is similar to that for RBI-SMART and is found in [57].

25.4 The ABEMML Algorithm

We make the same assumptions as in the previous section. The iterative step of the ABEMML algorithm is

$$x_j^{k+1} = \alpha_j^k v_j + (1 - \alpha_j^k) u_j, \tag{25.5}$$

where

$$\alpha_j^k = \gamma_j^k / d_j^k, \tag{25.6}$$

$$\gamma_j^k = (x_j^k - u_j) e_j^k, \tag{25.7}$$

$$\beta_j^k = (v_j - x_j^k) f_j^k, \tag{25.8}$$

$$d_j^k = \gamma_j^k + \beta_j^k, \tag{25.9}$$

$$e_j^k = \left(1 - \sum_{i \in B_n} A_{ij}\right) + \sum_{i \in B_n} A_{ij} \left(\frac{b_i - (Au)_i}{(Ax^k)_i - (Au)_i}\right), \tag{25.10}$$

and

$$f_j^k = \left(1 - \sum_{i \in B_n} A_{ij}\right) + \sum_{i \in B_n} A_{ij} \left(\frac{(Av)_i - b_i}{(Av)_i - (Ax^k)_i}\right). \tag{25.11}$$

We have the following theorem concerning the convergence of the ABEMML algorithm:

Theorem 25.2 *If there is a solution of the system $Ax = b$ that satisfies the constraints $u_j \leq x_j \leq v_j$ for each j, then, for any N and any choice of the blocks B_n, the ABEMML sequence converges to such a constrained solution of $Ax = b$. If there is no constrained solution of $Ax = b$, then, for $N = 1$, the ABEMML sequence converges to a constrained minimizer of*

$$KL(b - Au, Ax - Au) + KL(Av - b, Av - Ax).$$

The proof is similar to that for RBI-EMML and is to be found in [57]. In contrast to the ABMART theorem, this is all we can say about the limits of the ABEMML sequences.

Open Question: How does the limit of the ABEMML iterative sequence depend, in the consistent case, on the choice of blocks, and, in general, on the choice of x^0?

Chapter 26

Appendix: Block-Iterative Methods II

26.1 Chapter Summary

Image reconstruction problems in tomography are often formulated as statistical likelihood maximization problems in which the pixel values of the desired image play the role of parameters. Iterative algorithms based on cross-entropy minimization, such as the *expectation maximization maximum likelihood* (EMML) method and the *simultaneous multiplicative algebraic reconstruction technique* (SMART) can be used to solve such problems. Because the EMML and SMART are slow to converge for large amounts of data typical in imaging problems, acceleration of the algorithms using blocks of data or ordered subsets has become popular. There are a number of different ways to formulate these block-iterative versions of EMML and SMART, involving the choice of certain normalization and

regularization parameters. These methods are not faster merely because they are block-iterative; the correct choice of the parameters is crucial. The purpose of this chapter is to discuss these different formulations in detail sufficient to reveal the precise roles played by the parameters and to guide the user in choosing them.

It is not obvious, nor, in fact, is it even true, that using block-iterative methods will accelerate convergence. To better understand the connection between the use of blocks and acceleration, we begin with a discussion of the ART algorithm and its simultaneous versions, the Landweber algorithm and more particularly, Cimmino's method.

26.2 The ART and its Simultaneous Versions

In this section we let $Ax = b$ denote any real system of I linear equations in J unknowns. For each $i = 1, ..., I$ denote by H_i the hyperplane associated with the ith equation, that is,

$$H_i = \{x | (Ax)_i = b_i\},$$

and P_i the orthogonal projection operator onto H_i, that is, for every vector z, $P_i z$ is the vector in H_i closest to z. We can write $P_i z$ explicitly; we have

$$P_i z = z + (b_i - (Az)_i)a^i,$$

where a^i is the ith column of the matrix A^T, which we shall assume has been normalized to have $||a^i||_2 = 1$.

26.2.1 The ART

For $k = 0, 1, ...$ and $i = i(k) = k(\text{mod}\, I) + 1$, the ART iterative step is

$$x^{k+1} = P_i x^k = x^k + (b_i - (Ax^k)_i)a^i.$$

The ART operates by projecting the current vector onto the next hyperplane and cycling through the hyperplanes repeatedly. The ART uses only one equation at each step of the iteration.

Suppose that $\hat{x}$ is a solution of $Ax = b$, so that $A\hat{x} = b$. Each step of the ART gets us closer to $\hat{x}$, as the following calculations show.

We begin by calculating $||\hat{x} - x^{k+1}||_2^2$. We use

$$||\hat{x} - x^{k+1}||_2^2 = \langle \hat{x} - x^{k+1}, \hat{x} - x^{k+1} \rangle$$

and the definition of x^{k+1} to get

$$||\hat{x}-x^{k+1}||_2^2 = ||\hat{x}-x^k||_2^2 - 2\langle \hat{x}-x^k, (b_i-(Ax^k)_i)a^i\rangle + \langle (b_i-(Ax^k)_i)a^i, (b_i-(Ax^k)_i)a^i\rangle$$

$$= ||\hat{x}-x^k||_2^2 - 2(b_i-(Ax^k)_i)\langle \hat{x}-x^k, a^i\rangle + (b_i-(Ax^k)_i)^2$$

$$= ||\hat{x}-x^k||_2^2 - 2(b_i-(Ax^k)_i)^2 + (b_i-(Ax^k)_i)^2 = ||\hat{x}-x^k||_2^2 - (b_i-(Ax^k)_i)^2.$$

Therefore, we find that

$$||\hat{x}-x^k||_2^2 - ||\hat{x}-x^{k+1}||_2^2 = (b_i-(Ax^k)_i)^2. \tag{26.1}$$

Consequently, we know that

$$||\hat{x}-x^k||_2^2 \geq ||\hat{x}-x^{k+1}||_2^2.$$

It will help us later to know that

$$||\hat{x}-x^0||_2^2 - ||\hat{x}-x^I||_2^2 = \sum_{i=1}^{I}(b_i-(Ax^{i-1})_i)^2. \tag{26.2}$$

This measures how much closer to $\hat{x}$ we are after we have used all the equations one time.

There is one other consideration concerning the ART. From Equation (26.2) we see that it is helpful to have the quantities $(b_i-(Ax^{i-1})_i)^2$ large; if the equations are ordered in such a way that these quantities are not large, then the ART will not converge as quickly as it may otherwise do. This can easily happen if the equations correspond to discrete line integrals through the object and the lines are ordered so that each line is close to the previous line. Ordering the lines randomly, or in any way that avoids unfortunate ordering, greatly improves convergence speed [166].

Relaxation also helps to speed up the convergence of ART [238]. A relaxed version of ART has the following iterative step:

$$x^{k+1} = x^k + \beta(b_i-(Ax^k)_i)a^i,$$

where $0 < \beta \leq 1$.

26.2.2 The Landweber and Cimmino Algorithms

As we just saw, the ART uses one equation at a time and, at each step of the iteration, projects orthogonally onto the hyperplane associated with the next equation. A *simultaneous* version of ART, Cimmino's method, uses all the equations at each step, projecting orthogonally onto all the hyperplanes and averaging the result.

26.2.2.1 Cimmino's Algorithm:

The iterative step of Cimmino's algorithm is

$$x^{k+1} = x^k + \frac{1}{I}\sum_{i=1}^{I}(b_i - (Ax^k)_i)a^i = x^k + \frac{1}{I}A^T(b - Ax^k), \tag{26.3}$$

where, as previously, we assume that $||a^i||_2 = 1$ for all i.

26.2.2.2 Landweber's Algorithm:

A more general iterative algorithm is the Landweber algorithm, with the iterative step

$$x^{k+1} = x^k + \gamma A^T(b - Ax^k); \tag{26.4}$$

for convergence of this algorithm we need $0 \leq \gamma \leq 2/\rho(A^TA)$, where $\rho(A^TA)$ denotes the largest eigenvalue of the matrix A^TA. Since $||a^i||_2 = 1$ for all i, it follows that the trace of the matrix AA^T is I, which is also the trace of the matrix A^TA; since the trace of A^TA is also the sum of the eigenvalues of A^TA, it follows that the choice of $\gamma = \frac{1}{I}$ in Cimmino's method is acceptable.

Theorem 26.1 *For any γ in the interval $(0, \frac{2}{\rho(A^TA)})$, and any starting vector x^0, the Landweber iterative sequence given by Equation (26.4) converges to the minimizer of the function $f(x) = \frac{1}{2}\|Ax - b\|_2^2$ for which $\|x - x^0\|_2$ is minimized.*

Proof: Let $\hat{x}$ be a minimizer of $f(x)$. We calculate how much closer to $\hat{x}$ we get as we take one step of the Landweber iteration. We have

$$||\hat{x} - x^k||_2^2 - ||\hat{x} - x^{k+1}||_2^2$$

$$= 2\gamma\langle\hat{x} - x^k, A^T(b - Ax^k)\rangle - \gamma^2\langle A^T(b - Ax^k), A^T(b - Ax^k)\rangle. \tag{26.5}$$

From Theorem 8.2, we know that $b = A\hat{x} + \hat{w}$, where

$$\|b - A\hat{x}\|_2 \leq \|b - Ax\|_2,$$

for all x, and $A^T\hat{w} = 0$. Since every occurrence of b in Equation (26.5) is preceded by A^T, we can replace all those b with $A\hat{x}$. This gives us

$$||\hat{x} - x^k||_2^2 - ||\hat{x} - x^{k+1}||_2^2 = 2\gamma\langle\hat{x} - x^k, A^T(A\hat{x} - Ax^k)\rangle$$

$$-\gamma^2\langle A^T(A\hat{x} - Ax^k), A^T(A\hat{x} - Ax^k)\rangle. \tag{26.6}$$

From our earlier discussion of eigenvectors and eigenvalues, we know that, for any matrix B, we have

$$||Bx||_2^2 \leq \rho(B^T B)||x||_2^2.$$

Therefore,

$$\langle A^T(A\hat{x}-Ax^k), A^T(A\hat{x}-Ax^k)\rangle = ||A^T(A\hat{x}-Ax^k)||_2^2 \leq \rho(A^T A)||A\hat{x}-Ax^k||_2^2.$$

Using

$$\langle \hat{x}-x^k, A^T(A\hat{x}-Ax^k)\rangle = \langle A(\hat{x}-x^k), A\hat{x}-Ax^k\rangle = \langle A\hat{x}-Ax^k, A\hat{x}-Ax^k\rangle = ||A\hat{x}-Ax^k||_2^2,$$

we find that

$$||\hat{x}-x^k||_2^2 - ||\hat{x}-x^{k+1}||_2^2 \geq (2\gamma - \gamma^2\rho(A^T A))||A\hat{x}-Ax^k||_2^2.$$

For $0 < \gamma < \frac{2}{\rho(A^T A)}$ the sequence $\{||\hat{x}-x^k||_2^2\}$ is decreasing.
We can draw several conclusions from this inequality:

- 1. the sequence $\{\|\hat{x}-x^k\|_2\}$ is decreasing;
- 2. the sequence $\{\|A\hat{x}-Ax^k\|_2\}$ converges to zero;
- 3. the sequence $\{x^k\}$ is bounded;
- 4. there is a subsequence $\{x^{k_n}\}$ converging to some vector x^*;
- 5. $Ax^* = A\hat{x}$;
- 6. we can replace $\hat{x}$ with x^*, and so the sequence $\{\|x^*-x^k\|_2\}$ is decreasing;
- 7. since the subsequence $\{\|x^*-x^{k_n}\|_2\}$ converges to zero, so does the entire sequence $\{\|x^*-x^k\|_2\}$.

We can then say that the sequence $\{x^k\}$ converges to x^*, which is a minimizer of $f(x)$. Now we want to show that x^* is the minimizer closest to x^0.

We see from Equation (26.6) that the right side of the equation depends on $A\hat{x}$, but not directly on $\hat{x}$ itself. Therefore, the same is true of the difference

$$\|\hat{x}-x^0\|_2^2 - \|\hat{x}-x^*\|_2^2.$$

Therefore, minimizing $\|\hat{x}-x^0\|_2$ over all $\hat{x}$ that minimize $f(x)$ is equivalent to minimizing $\|\hat{x}-x^*\|_2$ over all $\hat{x}$ minimizing $f(x)$. But the answer to the latter problem is obviously $\hat{x} = x^*$. Therefore, x^* is the minimizer of $f(x)$ closest to x^0. ∎

If we take $\gamma = \frac{1}{\rho(A^TA)}$ we have

$$||\hat{x} - x^k||_2^2 - ||\hat{x} - x^{k+1}||_2^2 \geq \frac{1}{\rho(A^TA)}||A\hat{x} - Ax^k||_2^2. \tag{26.7}$$

In the case of Cimmino's method, we have $\gamma = \frac{1}{I}$, so that

$$||\hat{x} - x^k||_2^2 - ||\hat{x} - x^{k+1}||_2^2 \geq \frac{1}{I}||b - Ax^k||_2^2. \tag{26.8}$$

Using Equation (26.2) and the inequality in (26.8), we can make a rough comparison between ART and Cimmino's method, with respect to how much closer to $\hat{x}$ we get as we pass through all the equations one time. The two quantities

$$\sum_{i=1}^{I}(b_i - (Ax^{i-1})_i)^2$$

from Equation (26.2) and

$$||b - Ax^k||_2^2$$

from the inequality in (26.8) are comparable, in that both sums are over $i = 1, ..., I$, even though what is being summed is not the same in both cases. In image reconstruction I is quite large and the most important thing in such comparisons is the range of the summation index, so long as what is being summed is roughly comparable. However, notice that in the inequality in (26.8) the right side also has a factor of $\frac{1}{I}$. This tells us that, roughly speaking, one pass through all the equations using ART improves the squared distance to $\hat{x}$ by a factor of I, compared to using all the equations in one step of Cimmino's method, even though the amount of calculation is about the same.

Because the Landweber algorithm permits other choices for the parameter γ, there is hope that we may obtain better results with $\gamma \neq \frac{1}{I}$. The inequality

$$0 < \gamma < \frac{2}{\rho(A^TA)}$$

suggests using $\gamma = \frac{1}{\rho(A^TA)}$, which means that it would help to have a decent estimate of $\rho(A^TA)$; the estimate used in Cimmino's method is $\rho(A^TA) = I$, which is usually much too large. As a result, the choice of $\gamma = \frac{1}{I}$ means that we are taking unnecessarily small steps at each iteration. A smaller upper bound for $\rho(A^TA)$ would allow us to take bigger steps each time, and therefore, getting close to $\hat{x}$ sooner.

In many image processing applications, such as tomography, the matrix A is *sparse*, which means that most of the entries of A are zero. In the tomography problems for example, the number of non-zero entries of A is usually on the order of $\sqrt{J}$; since I and J are usually roughly comparable,

this means that A has about $\sqrt{I}$ non-zero entries. In the appendix on matrix theory we obtain an upper bound estimate for $\rho(A^TA)$ that is particularly useful when A is sparse. Suppose that all the rows of A have length one. Let s be the largest number of non-zero entries in any column of A. Then $\rho(A^TA)$ does not exceed s. Notice that this estimate does not require us to calculate the matrix A^TA and makes use of the sparse nature of A; the matrix A^TA need not be sparse, and would be time-consuming to calculate in practice, anyway.

If, for the sparse cases, we take $\rho(A^TA)$ to be approximately $\sqrt{I}$, and choose $\gamma = \frac{1}{\sqrt{I}}$, we find that we have replaced the factor $\frac{1}{I}$ in the inequality (26.8) with the much larger factor $\frac{1}{\sqrt{I}}$, which then improves the rate of convergence. However, the ART is still faster by, roughly, a factor of $\sqrt{I}$.

26.2.3 Block-Iterative ART

The ART uses only one equation at a time, while the Landweber algorithm uses all the equations at each step of the iteration. It is sometimes convenient to take a middle course, and use some, but not all, equations at each step of the iteration. The collection of equations to be used together constitute a *block*. Such methods are called *block-iterative* or *ordered-subset* methods. Generally speaking, when unfortunate ordering of the blocks and selection of equations within each block are avoided, and the parameters well chosen, these block-iterative methods converge faster than the Cimmino algorithm by roughly a factor of the number of blocks.

We turn now to the iterative algorithms that are based on the KL distance. For these algorithms as well, we find that using block-iterative methods and choosing the parameters carefully, we can improve convergence by roughly the number of blocks, with respect to the simultaneous EMML and SMART methods.

26.3 Overview of KL-based methods

The algorithms we discuss here have interesting histories, which we sketch in this section.

26.3.1 The SMART and its variants

Like the ART, the MART has a simultaneous version, called the SMART. Like MART, SMART applies only to nonnegative systems of equations $Ax = b$. Unlike MART, SMART is a simultaneous algorithm that uses

all equations in each step of the iteration. The SMART was discovered in 1972, independently, by Darroch and Ratcliff, working in statistics, [107] and by Schmidlin [233] in medical imaging; neither work makes reference to MART. Darroch and Ratcliff do consider block-iterative versions of their algorithm, in which only some of the equations are used at each step, but their convergence proof involves unnecessary restrictions on the system matrix. Censor and Segman [89] seem to be the first to present the SMART and its block-iterative variants explicitly as generalizations of MART.

26.3.2 The EMML and its variants

The *expectation maximization maximum likelihood* (EMML) method turns out to be closely related to the SMART, although it has quite a different history. The EMML algorithm we discuss here is actually a special case of a more general approach to likelihood maximization, usually called the EM algorithm [109]; the book by McLachnan and Krishnan [202] is a good source for the history of this more general algorithm.

It was noticed by Rockmore and Macovski [230] that the image reconstruction problems posed by medical tomography could be formulated as statistical parameter estimation problems. Following up on this idea, Shepp and Vardi [237] suggested the use of the EM algorithm for solving the reconstruction problem in emission tomography. In [189], Lange and Carson presented an EM-type iterative method for transmission tomographic image reconstruction, and pointed out a gap in the convergence proof given in [237] for the emission case. In [258], Vardi, Shepp and Kaufman repaired the earlier proof, relying on techniques due to Csiszár and Tusnády [104]. In [190] Lange, Bahn and Little improved the transmission and emission algorithms, by including regularization to reduce the effects of noise. The question of uniqueness of the solution in the inconsistent case was resolved in [51].

The MART and SMART were initially designed to apply to consistent systems of equations. Darroch and Ratcliff did not consider what happens in the inconsistent case, in which the system of equations has no non-negative solutions; this issue was resolved in [51], where it was shown that the SMART converges to a non-negative minimizer of the Kullback-Leibler distance $KL(Ax, b)$. The EMML, as a statistical parameter estimation technique, was not originally thought to be connected to any system of linear equations. In [51] it was shown that the EMML leads to a non-negative minimizer of the Kullback-Leibler distance $KL(b, Ax)$, thereby exhibiting a close connection between the SMART and the EMML methods. Consequently, when the non-negative system of linear equations $Ax = b$ has a non-negative solution, the EMML converges to such a solution.

26.3.3 Block-iterative Versions of SMART and EMML

As we have seen, Darroch and Ratcliff included what are now called block-iterative versions of SMART in their original paper [107]. Censor and Segman [89] viewed SMART and its block-iterative versions as natural extension of the MART. Consequently, block-iterative variants of SMART have been around for some time. The story with the EMML is quite different.

The paper of Holte, Schmidlin, *et al.* [170] compares the performance of Schmidlin's method of [233] with the EMML algorithm. Almost as an aside, they notice the accelerating effect of what they call *projection interleaving*, that is, the use of blocks. This paper contains no explicit formulas, however, and presents no theory, so one can only make educated guesses as to the precise iterative methods employed. Somewhat later, Hudson, Hutton and Larkin [171, 172] observed that the EMML can be significantly accelerated if, at each step, one employs only some of the data. They referred to this approach as the *ordered subset* EM method (OSEM). They gave a proof of convergence of the OSEM, for the consistent case. The proof relied on a fairly restrictive relationship between the matrix A and the choice of blocks, called *subset balance*. In [54] a revised version of the OSEM, called the *rescaled block-iterative* EMML (RBI-EMML), was shown to converge, in the consistent case, regardless of the choice of blocks.

26.3.4 Basic assumptions

Methods based on cross-entropy, such as the MART, SMART, EMML and all block-iterative versions of these algorithms apply to nonnegative systems that we denote by $Ax = b$, where b is a vector of positive entries, A is a matrix with entries $A_{ij} \geq 0$ such that for each j the sum $s_j = \sum_{i=1}^{I} A_{ij}$ is positive and we seek a solution x with nonnegative entries. If no nonnegative x satisfies $b = Ax$ we say the system is *inconsistent*.

Simultaneous iterative algorithms employ all of the equations at each step of the iteration; block-iterative methods do not. For the latter methods we assume that the index set $\{i = 1, ..., I\}$ is the (not necessarily disjoint) union of the N sets or *blocks* B_n, $n = 1, ..., N$. We shall require that $s_{nj} = \sum_{i \in B_n} A_{ij} > 0$ for each n and each j. Block-iterative methods like ART and MART for which each block consists of precisely one element are called *row-action* or *sequential* methods. We begin our discussion with the SMART and the EMML method.

26.4 The SMART and the EMML method

Both the SMART and the EMML method provide a solution of $b = Ax$ when such exist and (distinct) approximate solutions in the inconsistent case. The SMART algorithm is the following:

Algorithm 26.1 (SMART) *Let x^0 be an arbitrary positive vector. For $k = 0, 1, ...$ let*

$$x_j^{k+1} = x_j^k \exp\Big(s_j^{-1} \sum_{i=1}^{I} A_{ij} \log \frac{b_i}{(Ax^k)_i}\Big). \tag{26.9}$$

The exponential and logarithm in the SMART iterative step are computationally expensive. The EMML method is similar to the SMART, but somewhat less costly to compute.

Algorithm 26.2 (EMML) *Let x^0 be an arbitrary positive vector. For $k = 0, 1, ...$ let*

$$x_j^{k+1} = x_j^k s_j^{-1} \sum_{i=1}^{I} A_{ij} \frac{b_i}{(Ax^k)_i}. \tag{26.10}$$

The main results concerning the SMART are given by the following theorem.

Theorem 26.2 *In the consistent case the SMART converges to the unique nonnegative solution of $b = Ax$ for which the distance $\sum_{j=1}^{J} s_j KL(x_j, x_j^0)$ is minimized. In the inconsistent case it converges to the unique nonnegative minimizer of the distance $KL(Ax, y)$ for which $\sum_{j=1}^{J} s_j KL(x_j, x_j^0)$ is minimized; if A and every matrix derived from A by deleting columns has full rank then there is a unique nonnegative minimizer of $KL(Ax, y)$ and at most $I - 1$ of its entries are nonzero.*

For the EMML method the main results are the following.

Theorem 26.3 *In the consistent case the EMML algorithm converges to nonnegative solution of $b = Ax$. In the inconsistent case it converges to a nonnegative minimizer of the distance $KL(y, Ax)$; if A and every matrix derived from A by deleting columns has full rank then there is a unique nonnegative minimizer of $KL(y, Ax)$ and at most $I - 1$ of its entries are nonzero.*

In the consistent case there may be multiple nonnegative solutions and the one obtained by the EMML algorithm will depend on the starting vector x^0; how it depends on x^0 is an open question.

These theorems are special cases of more general results on block-iterative methods that we shall prove later in this chapter.

Both the EMML and SMART are related to likelihood maximization. Minimizing the function $KL(y, Ax)$ is equivalent to maximizing the likelihood when the b_i are taken to be measurements of independent Poisson random variables having means $(Ax)_i$. The entries of x are the parameters to be determined. This situation arises in emission tomography. So the EMML is a likelihood maximizer, as its name suggests.

The connection between SMART and likelihood maximization is a bit more convoluted. Suppose that $s_j = 1$ for each j. The solution of $b = Ax$ for which $KL(x, x^0)$ is minimized necessarily has the form

$$x_j = x_j^0 \exp\Big(\sum_{i=1}^{I} A_{ij}\lambda_i\Big) \tag{26.11}$$

for some vector λ with entries λ_i. This *log linear* form also arises in transmission tomography, where it is natural to assume that $s_j = 1$ for each j and $\lambda_i \leq 0$ for each i. We have the following lemma that helps to connect the SMART algorithm with the transmission tomography problem:

Lemma 26.1 *Minimizing $KL(d, x)$ over x as in Equation (26.11) is equivalent to minimizing $KL(x, x^0)$, subject to $Ax = Ad$.*

The solution to the latter problem can be obtained using the SMART.

With $x_+ = \sum_{j=1}^{J} x_j$ the vector A with entries $p_j = x_j/x_+$ is a probability vector. Let $d = (d_1, ..., d_J)^T$ be a vector whose entries are nonnegative integers, with $K = \sum_{j=1}^{J} d_j$. Suppose that, for each j, p_j is the probability of index j and d_j is the number of times index j was chosen in K trials. The likelihood function of the parameters λ_i is

$$L(\lambda) = \prod_{j=1}^{J} p_j^{d_j} \tag{26.12}$$

so that the log-likelihood function is

$$LL(\lambda) = \sum_{j=1}^{J} d_j \log p_j. \tag{26.13}$$

Since A is a probability vector, maximizing $L(\lambda)$ is equivalent to minimizing $KL(d, p)$ with respect to λ, which, according to the lemma above, can be solved using SMART. In fact, since all of the block-iterative versions of SMART have the same limit whenever they have the same starting vector, any of these methods can be used to solve this maximum likelihood problem. In the case of transmission tomography the λ_i must be non-positive,

so if SMART is to be used, some modification is needed to obtain such a solution.

Those who have used the SMART or the EMML on sizable problems have certainly noticed that they are both slow to converge. An important issue, therefore, is how to accelerate convergence. One popular method is through the use of *block-iterative* (or *ordered subset*) methods.

26.5 Ordered-Subset Versions

To illustrate block-iterative methods and to motivate our subsequent discussion we consider now the *ordered subset* EM algorithm (OSEM), which is a popular technique in some areas of medical imaging, as well as an analogous version of SMART, which we shall call here the OSSMART. The OSEM is now used quite frequently in tomographic image reconstruction, where it is acknowledged to produce usable images significantly faster then EMML. From a theoretical perspective both OSEM and OSSMART are incorrect. How to correct them is the subject of much that follows here.

The idea behind the OSEM (OSSMART) is simple: the iteration looks very much like the EMML (SMART), but at each step of the iteration the summations are taken only over the current block. The blocks are processed cyclically.

The OSEM iteration is the following: for $k = 0, 1, ...$ and $n = k(\mathrm{mod}\, N) + 1$, having found x^k let

OSEM:

$$x_j^{k+1} = x_j^k s_{nj}^{-1} \sum_{i \in B_n} A_{ij} \frac{b_i}{(Ax^k)_i}. \tag{26.14}$$

The OSSMART has the following iterative step:

OSSMART

$$x_j^{k+1} = x_j^k \exp\Big(s_{nj}^{-1} \sum_{i \in B_n} A_{ij} \log \frac{b_i}{(Ax^k)_i}\Big). \tag{26.15}$$

In general we do not expect block-iterative algorithms to converge in the inconsistent case, but to exhibit *subsequential convergence* to a *limit cycle*, as we shall discuss later. We do, however, want them to converge to a solution in the consistent case; the OSEM and OSSMART fail to do this except when the matrix A and the set of blocks $\{B_n, n = 1, ..., N\}$ satisfy the condition known as *subset balance*, which means that the sums s_{nj}

depend only on j and not on n. While this may be approximately valid in some special cases, it is overly restrictive, eliminating, for example, almost every set of blocks whose cardinalities are not all the same. When the OSEM does well in practice in medical imaging it is probably because the N is not large and only a few iterations are carried out.

The experience with the OSEM was encouraging, however, and strongly suggested that an equally fast, but mathematically correct, block-iterative version of EMML was to be had; this is the *rescaled block-iterative* EMML (RBI-EMML). Both RBI-EMML and an analogous corrected version of OSSMART, the RBI-SMART, provide fast convergence to a solution in the consistent case, for any choice of blocks.

26.6 The RBI-SMART

We turn next to the block-iterative versions of the SMART, which we shall denote BI-SMART. These methods were known prior to the discovery of RBI-EMML and played an important role in that discovery; the importance of rescaling for acceleration was apparently not appreciated, however.

We start by considering a formulation of BI-SMART that is general enough to include all of the variants we wish to discuss. As we shall see, this formulation is too general and will need to be restricted in certain ways to obtain convergence. Let the iterative step be

$$x_j^{k+1} = x_j^k \exp\Big(\beta_{nj} \sum_{i \in B_n} \alpha_{ni} A_{ij} \log\Big(\frac{b_i}{(Ax^k)_i}\Big)\Big), \tag{26.16}$$

for $j = 1, 2, ..., J$, $n = k(\mathrm{mod}\, N) + 1$ and β_{nj} and α_{ni} positive. As we shall see, our convergence proof will require that β_{nj} be separable, that is, $\beta_{nj} = \gamma_j \delta_n$ for each j and n and that

$$\gamma_j \delta_n \sigma_{nj} \leq 1, \tag{26.17}$$

for $\sigma_{nj} = \sum_{i \in B_n} \alpha_{ni} A_{ij}$. With these conditions satisfied we have the following result.

Theorem 26.4 *Suppose that we are in the consistent case, in which the system $Ax = b$ has non-negative solutions. For any positive vector x^0 and any collection of blocks $\{B_n, n = 1, ..., N\}$ the sequence $\{x^k\}$ given by Equation (26.16) converges to the unique solution of $b = Ax$ for which the weighted cross-entropy $\sum_{j=1}^J \gamma_j^{-1} KL(x_j, x_j^0)$ is minimized.*

The inequality in the following lemma is the basis for the convergence proof.

Lemma 26.2 *Let $b = Ax$ for some nonnegative x. Then for $\{x^k\}$ as in Equation (26.16) we have*

$$\sum_{j=1}^{J} \gamma_j^{-1} KL(x_j, x_j^k) - \sum_{j=1}^{J} \gamma_j^{-1} KL(x_j, x_j^{k+1}) \geq \tag{26.18}$$

$$\delta_n \sum_{i \in B_n} \alpha_{ni} KL(b_i, (Ax^k)_i). \tag{26.19}$$

Proof: First note that

$$x_j^{k+1} = x_j^k \exp\Big(\gamma_j \delta_n \sum_{i \in B_n} \alpha_{ni} A_{ij} \log\Big(\frac{b_i}{(Ax^k)_i}\Big)\Big), \tag{26.20}$$

and

$$\exp\Big(\gamma_j \delta_n \sum_{i \in B_n} \alpha_{ni} A_{ij} \log\Big(\frac{b_i}{(Ax^k)_i}\Big)\Big) \tag{26.21}$$

can be written as

$$\exp\Big((1 - \gamma_j \delta_n \sigma_{nj}) \log 1 + \gamma_j \delta_n \sum_{i \in B_n} \alpha_{ni} A_{ij} \log\Big(\frac{b_i}{(Ax^k)_i}\Big)\Big), \tag{26.22}$$

which, by the convexity of the exponential function, is not greater than

$$(1 - \gamma_j \delta_n \sigma_{nj}) + \gamma_j \delta_n \sum_{i \in B_n} \alpha_{ni} A_{ij} \frac{b_i}{(Ax^k)_i}. \tag{26.23}$$

It follows that

$$\sum_{j=1}^{J} \gamma_j^{-1} (x_j^k - x_j^{k+1}) \geq \delta_n \sum_{i \in B_n} \alpha_{ni} ((Ax^k)_i - b_i). \tag{26.24}$$

We also have

$$\log(x_j^{k+1}/x_j^k) = \gamma_j \delta_n \sum_{i \in B_n} \alpha_{ni} A_{ij} \log \frac{b_i}{(Ax^k)_i}. \tag{26.25}$$

Therefore

$$\sum_{j=1}^{J} \gamma_j^{-1} KL(x_j, x_j^k) - \sum_{j=1}^{J} \gamma_j^{-1} KL(x_j, x_j^{k+1}) \tag{26.26}$$

$$= \sum_{j=1}^{J} \gamma_j^{-1}(x_j \log(x_j^{k+1}/x_j^k) + x_j^k - x_j^{k+1}) \tag{26.27}$$

$$= \sum_{j=1}^{J} x_j \delta_n \sum_{i \in B_n} \alpha_{ni} A_{ij} \log \frac{b_i}{(Ax^k)_i} + \sum_{j=1}^{J} \gamma_j^{-1}(x_j^k - x_j^{k+1}) \tag{26.28}$$

$$= \delta_n \sum_{i \in B_n} \alpha_{ni} (\sum_{j=1}^{J} x_j A_{ij}) \log \frac{b_i}{(Ax^k)_i} + \sum_{j=1}^{J} \gamma_j^{-1}(x_j^k - x_j^{k+1}) \tag{26.29}$$

$$\geq \delta_n \Big(\sum_{i \in B_n} \alpha_{ni} (b_i \log \frac{b_i}{(Ax^k)_i} + (Ax^k)_i - b_i) \Big) = \delta_n \sum_{i \in B_n} \alpha_{ni} KL(b_i, (Ax^k)_i). \tag{26.30}$$

This completes the proof of the lemma. ∎

From the inequality (26.19) we conclude that the sequence

$$\{\sum_{j=1}^{J} \gamma_j^{-1} KL(x_j, x_j^k)\} \tag{26.31}$$

is decreasing, that $\{x^k\}$ is therefore bounded and the sequence

$$\{\sum_{i \in B_n} \alpha_{ni} KL(b_i, (Ax^k)_i)\} \tag{26.32}$$

is converging to zero. Let x^* be any cluster point of the sequence $\{x^k\}$. Then it is not difficult to show that $b = Ax^*$. Replacing x with x^* we have that the sequence $\{\sum_{j=1}^{J} \gamma_j^{-1} KL(x_j^*, x_j^k)\}$ is decreasing; since a subsequence converges to zero, so does the whole sequence. Therefore x^* is the limit of the sequence $\{x^k\}$. This proves that the algorithm produces a solution of $b = Ax$. To conclude further that the solution is the one for which the quantity $\sum_{j=1}^{J} \gamma_j^{-1} KL(x_j, x_j^0)$ is minimized requires further work to replace the inequality (26.19) with an equation in which the right side is independent of the particular solution x chosen; see the final section of this chapter for the details.

We see from the theorem that how we select the γ_j is determined by how we wish to weight the terms in the sum $\sum_{j=1}^{J} \gamma_j^{-1} KL(x_j, x_j^0)$. In some cases we want to minimize the cross-entropy $KL(x, x^0)$ subject to $b = Ax$; in this case we would select $\gamma_j = 1$. In other cases we may have some prior knowledge as to the relative sizes of the x_j and wish to emphasize

the smaller values more; then we may choose γ_j proportional to our prior estimate of the size of x_j. Having selected the γ_j, we see from the inequality (26.19) that convergence will be accelerated if we select δ_n as large as permitted by the condition $\gamma_j \delta_n \sigma_{nj} \leq 1$. This suggests that we take

$$\delta_n = 1/\min\{\sigma_{nj}\gamma_j,\, j = 1, ..., J\}. \tag{26.33}$$

The *rescaled* BI-SMART (RBI-SMART) as presented in [53, 55, 56] uses this choice, but with $\alpha_{ni} = 1$ for each n and i. For each $n = 1, ..., N$ let

$$m_n = \max\{s_{nj} s_j^{-1} | j = 1, ..., J\}. \tag{26.34}$$

The original RBI-SMART is as follows:

Algorithm 26.3 (RBI-SMART) *Let x^0 be an arbitrary positive vector. For $k = 0, 1, ...$, let $n = k(\mathrm{mod}\, N) + 1$. Then let*

$$x_j^{k+1} = x_j^k \exp\Big(m_n^{-1} s_j^{-1} \sum_{i \in B_n} A_{ij} \log\Big(\frac{b_i}{(Ax^k)_i}\Big)\Big). \tag{26.35}$$

Notice that Equation (26.35) can be written as

$$\log x_j^{k+1} = (1 - m_n^{-1} s_j^{-1} s_{nj}) \log x_j^k + m_n^{-1} s_j^{-1} \sum_{i \in B_n} A_{ij} \log\Big(x_j^k \frac{b_i}{(Ax^k)_i}\Big), \tag{26.36}$$

from which we see that x_j^{k+1} is a weighted geometric mean of x_j^k and the terms

$$(Q_i x^k)_j = x_j^k \Big(\frac{b_i}{(Ax^k)_i}\Big),$$

for $i \in B_n$. This will be helpful in deriving block-iterative versions of the EMML algorithm. The vectors $Q_i(x^k)$ are sometimes called weighted KL projections.

Let's look now at some of the other choices for these parameters that have been considered in the literature.

First, we notice that the OSSMART does not generally satisfy the requirements, since in (26.15) the choices are $\alpha_{ni} = 1$ and $\beta_{nj} = s_{nj}^{-1}$; the only times this is acceptable is if the s_{nj} are separable; that is, $s_{nj} = r_j t_n$ for some r_j and t_n. This is slightly more general than the condition of subset balance and is sufficient for convergence of OSSMART.

In [89] Censor and Segman make the choices $\beta_{nj} = 1$ and $\alpha_{ni} > 0$ such that $\sigma_{nj} \leq 1$ for all n and j. In those cases in which σ_{nj} is much less than 1 for each n and j their iterative scheme is probably excessively relaxed; it is hard to see how one might improve the rate of convergence by altering only the weights α_{ni}, however. Limiting the choice to $\gamma_j \delta_n = 1$ reduces our ability to accelerate this algorithm.

The original SMART in Equation (26.9) uses $N = 1$, $\gamma_j = s_j^{-1}$ and $\alpha_{ni} = \alpha_i = 1$. Clearly the inequality (26.17) is satisfied; in fact it becomes an equality now.

For the row-action version of SMART, the *multiplicative* ART (MART), due to Gordon, Bender and Herman [151], we take $N = I$ and $B_n = B_i = \{i\}$ for $i = 1, ..., I$. The MART has the iterative

$$x_j^{k+1} = x_j^k \Big(\frac{b_i}{(Ax^k)_i} \Big)^{m_i^{-1} A_{ij}}, \tag{26.37}$$

for $j = 1, 2, ..., J$, $i = k(\mathrm{mod}\, I) + 1$ and $m_i > 0$ chosen so that $m_i^{-1} A_{ij} \leq 1$ for all j. The smaller m_i is the faster the convergence, so a good choice is $m_i = \max\{A_{ij}|, j = 1, ..., J\}$. Although this particular choice for m_i is not explicitly mentioned in the various discussions of MART I have seen, it was used in implementations of MART from the beginning [165].

Darroch and Ratcliff included a discussion of a block-iterative version of SMART in their 1972 paper [107]. Close inspection of their version reveals that they require that $s_{nj} = \sum_{i \in B_n} A_{ij} = 1$ for all j. Since this is unlikely to be the case initially, we might try to rescale the equations or unknowns to obtain this condition. However, unless $s_{nj} = \sum_{i \in B_n} A_{ij}$ depends only on j and not on n, which is the *subset balance* property used in [172], we cannot redefine the unknowns in a way that is independent of n.

The MART fails to converge in the inconsistent case. What is always observed, but for which no proof exists, is that, for each fixed $i = 1, 2, ..., I$, as $m \to +\infty$, the MART subsequences $\{x^{mI+i}\}$ converge to separate limit vectors, say $x^{\infty,i}$. This *limit cycle* LC $= \{x^{\infty,i} | i = 1, ..., I\}$ reduces to a single vector whenever there is a nonnegative solution of $b = Ax$. The greater the minimum value of $KL(Ax, b)$ the more distinct from one another the vectors of the limit cycle are. An analogous result is observed for BI-SMART.

26.7 The RBI-EMML

As we did with SMART, we consider now a formulation of BI-EMML that is general enough to include all of the variants we wish to discuss. Once again, the formulation is too general and will need to be restricted in certain ways to obtain convergence. Let the iterative step be

$$x_j^{k+1} = x_j^k (1 - \beta_{nj} \sigma_{nj}) + x_j^k \beta_{nj} \sum_{i \in B_n} \alpha_{ni} A_{ij} \frac{b_i}{(Ax^k)_i}, \tag{26.38}$$

for $j = 1, 2, ..., J$, $n = k(\mathrm{mod}\, N)+1$ and β_{nj} and α_{ni} positive. As in the case of BI-SMART, our convergence proof will require that β_{nj} be separable, that is,

$$\beta_{nj} = \gamma_j \delta_n \tag{26.39}$$

for each j and n and that the inequality (26.17) hold. With these conditions satisfied we have the following result.

Theorem 26.5 *Suppose that we are in the consistent case. For any positive vector x^0 and any collection of blocks $\{B_n, n = 1, ..., N\}$ the sequence $\{x^k\}$ given by Equation (26.16) converges to a nonnegative solution of $b = Ax$.*

When there are multiple nonnegative solutions of $b = Ax$ the solution obtained by BI-EMML will depend on the starting point x^0, but precisely how it depends on x^0 is an open question. Also, in contrast to the case of BI-SMART, the solution can depend on the particular choice of the blocks. The inequality in the following lemma is the basis for the convergence proof.

Lemma 26.3 *Let $Ax = b$ for some non-negative x. Then, for $\{x^k\}$ as in Equation (26.38), we have*

$$\sum_{j=1}^{J} \gamma_j^{-1} KL(x_j, x_j^k) - \sum_{j=1}^{J} \gamma_j^{-1} KL(x_j, x_j^{k+1}) \geq \tag{26.40}$$

$$\delta_n \sum_{i \in B_n} \alpha_{ni} KL(b_i, (Ax^k)_i). \tag{26.41}$$

Proof: From the iterative step

$$x_j^{k+1} = x_j^k (1 - \gamma_j \delta_n \sigma_{nj}) + x_j^k \gamma_j \delta_n \sum_{i \in B_n} \alpha_{ni} A_{ij} \frac{b_i}{(Ax^k)_i} \tag{26.42}$$

we have

$$\log(x_j^{k+1}/x_j^k) = \log\Big((1 - \gamma_j \delta_n \sigma_{nj}) + \gamma_j \delta_n \sum_{i \in B_n} \alpha_{ni} A_{ij} \frac{b_i}{(Ax^k)_i}\Big). \tag{26.43}$$

By the concavity of the logarithm we obtain the inequality

$$\log(x_j^{k+1}/x_j^k) \geq \Big((1 - \gamma_j \delta_n \sigma_{nj}) \log 1 + \gamma_j \delta_n \sum_{i \in B_n} \alpha_{ni} A_{ij} \log \frac{b_i}{(Ax^k)_i}\Big), \tag{26.44}$$

or

$$\log(x_j^{k+1}/x_j^k) \geq \gamma_j \delta_n \sum_{i \in B_n} \alpha_{ni} A_{ij} \log \frac{b_i}{(Ax^k)_i}. \tag{26.45}$$

Therefore

$$\sum_{j=1}^{J} \gamma_j^{-1} x_j \log(x_j^{k+1}/x_j^k) \geq \delta_n \sum_{i \in B_n} \alpha_{ni} (\sum_{j=1}^{J} x_j A_{ij}) \log \frac{b_i}{(Ax^k)_i}. \tag{26.46}$$

Note that it is at this step that we used the separability of the β_{nj}. Also

$$\sum_{j=1}^{J} \gamma_j^{-1} (x_j^{k+1} - x_j^k) = \delta_n \sum_{i \in B_n} ((Ax^k)_i - b_i). \tag{26.47}$$

This concludes the proof of the lemma. ∎

From the inequality in (26.41) we conclude, as we did in the BI-SMART case, that the sequence $\{\sum_{j=1}^{J} \gamma_j^{-1} KL(x_j, x_j^k)\}$ is decreasing, that $\{x^k\}$ is therefore bounded and the sequence $\{\sum_{i \in B_n} \alpha_{ni} KL(b_i, (Ax^k)_i)\}$ is converging to zero. Let x^* be any cluster point of the sequence $\{x^\}$. Then it is not difficult to show that $b = Ax^*$. Replacing x with x^* we have that the sequence $\{\sum_{j=1}^{J} \gamma_j^{-1} KL(x_j^*, x_j^k)\}$ is decreasing; since a subsequence converges to zero, so does the whole sequence. Therefore x^* is the limit of the sequence $\{x^k\}$. This proves that the algorithm produces a nonnegative solution of $b = Ax$. So far, we have been unable to replace the inequality in (26.41) with an equation in which the right side is independent of the particular solution x chosen.

Having selected the γ_j, we see from the inequality in (26.41) that convergence will be accelerated if we select δ_n as large as permitted by the condition $\gamma_j \delta_n \sigma_{nj} \leq 1$. This suggests that once again we take

$$\delta_n = 1/\min\{\sigma_{nj}\gamma_j,\, j = 1, ..., J\}. \tag{26.48}$$

The *rescaled* BI-EMML (RBI-EMML) as presented in [53, 55, 56] uses this choice, but with $\alpha_{ni} = 1$ for each n and i. The original motivation for the RBI-EMML came from consideration of Equation (26.36), replacing the geometric means with arithmetic means. This RBI-EMML is as follows:

Algorithm 26.4 (RBI-EMML) *Let x^0 be an arbitrary positive vector. For $k = 0, 1, ...$, let $n = k(\mathrm{mod}\, N) + 1$. Then let*

$$x_j^{k+1} = (1 - m_n^{-1} s_j^{-1} s_{nj}) x_j^k + m_n^{-1} s_j^{-1} x_j^k \sum_{i \in B_n} (A_{ij} \frac{b_i}{(Ax^k)_i}). \tag{26.49}$$

Let's look now at some of the other choices for these parameters that have been considered in the literature.

First, we notice that the OSEM does not generally satisfy the requirements, since in (26.14) the choices are $\alpha_{ni} = 1$ and $\beta_{nj} = s_{nj}^{-1}$; the only times this is acceptable is if the s_{nj} are separable; that is, $s_{nj} = r_j t_n$ for some r_j and t_n. This is slightly more general than the condition of subset balance and is sufficient for convergence of OSEM.

The original EMML in Equation (26.10) uses $N = 1$, $\gamma_j = s_j^{-1}$ and $\alpha_{ni} = \alpha_i = 1$. Clearly the inequality (26.17) is satisfied; in fact it becomes an equality now.

Notice that the calculations required to perform the BI-SMART are somewhat more complicated than those needed in BI-EMML. Because the MART converges rapidly in most cases there is considerable interest in the row-action version of EMML. It was clear from the outset that using the OSEM in a row-action mode does not work. We see from the formula for BI-EMML that the proper row-action version of EMML, which we call the EM-MART, is the following:

Algorithm 26.5 (EM-MART) *Let x^0 be an arbitrary positive vector and $i = k(\mathrm{mod}\, I) + 1$. Then let*

$$x_j^{k+1} = (1 - \delta_i \gamma_j \alpha_{ii} A_{ij}) x_j^k + \delta_i \gamma_j \alpha_{ii} x_j^k A_{ij} \frac{b_i}{(Ax^k)_i}, \tag{26.50}$$

with

$$\gamma_j \delta_i \alpha_{ii} A_{ij} \leq 1 \tag{26.51}$$

for all i and j.

The optimal choice would seem to be to take $\delta_i \alpha_{ii}$ as large as possible; that is, to select $\delta_i \alpha_{ii} = 1/\max\{\gamma_j A_{ij}, j = 1, ..., J\}$. With this choice the EM-MART is called the *rescaled* EM-MART (REM-MART).

The EM-MART fails to converge in the inconsistent case. What is always observed, but for which no proof exists, is that, for each fixed $i = 1, 2, ..., I$, as $m \to +\infty$, the EM-MART subsequences $\{x^{mI+i}\}$ converge to separate limit vectors, say $x^{\infty,i}$. This *limit cycle* $\mathrm{LC} = \{x^{\infty,i} | i = 1, ..., I\}$ reduces to a single vector whenever there is a nonnegative solution of $b = Ax$. The greater the minimum value of $KL(b, Ax)$ the more distinct from one another the vectors of the limit cycle are. An analogous result is observed for BI-EMML.

We must mention a method that closely resembles the REM-MART, the *row-action maximum likelihood algorithm* (RAMLA), which was discovered independently by Browne and De Pierro [33]. The RAMLA avoids the limit cycle in the inconsistent case by using strong underrelaxation involving a decreasing sequence of relaxation parameters λ_k. The RAMLA is the following:

Algorithm 26.6 (RAMLA) *Let x^0 be an arbitrary positive vector, and $n = k(\mathrm{mod}\, N) + 1$. Let the positive relaxation parameters λ_k be chosen to converge to zero and $\sum_{k=0}^{+\infty} \lambda_k = +\infty$. Then,*

$$x_j^{k+1} = (1 - \lambda_k \sum_{i \in B_n} A_{ij}) x_j^k + \lambda_k x_j^k \sum_{i \in B_n} A_{ij} \Big(\frac{b_i}{(Ax^k)_i} \Big), \tag{26.52}$$

26.8 RBI-SMART and Entropy Maximization

As we stated earlier, in the consistent case the sequence $\{x^k\}$ generated by the BI-SMART algorithm and given by Equation (26.20) converges to the unique solution of $b = Ax$ for which the distance $\sum_{j=1}^{J} \gamma_j^{-1} KL(x_j, x_j^0)$ is minimized. In this section we sketch the proof of this result as a sequence of lemmas, each of which is easily established.

Lemma 26.4 *For any nonnegative vectors a and b with $a_+ = \sum_{m=1}^{M} a_m$ and $b_+ = \sum_{m=1}^{M} b_m > 0$ we have*

$$KL(a, b) = KL(a_+, b_+) + KL(a_+, \frac{a_+}{b_+} b). \tag{26.53}$$

For nonnegative vectors x and z let

$$G_n(x, z) = \sum_{j=1}^{J} \gamma_j^{-1} KL(x_j, z_j) \tag{26.54}$$

$$+\delta_n \sum_{i \in B_n} \alpha_{ni} [KL((Ax)_i, b_i) - KL((Ax)_i, (Az)_i)]. \tag{26.55}$$

It follows from Equation 26.53 and the inequality

$$\gamma_j^{-1} - \delta_n \sigma_{nj} \geq 1 \tag{26.56}$$

that $G_n(x, z) \geq 0$ in all cases.

Lemma 26.5 *For every x we have*

$$G_n(x, x) = \delta_n \sum_{i \in B_n} \alpha_{ni} KL((Ax)_i, b_i) \tag{26.57}$$

so that

$$G_n(x, z) = G_n(x, x) + \sum_{j=1}^{J} \gamma_j^{-1} KL(x_j, z_j) \tag{26.58}$$

$$-\delta_n \sum_{i \in B_n} \alpha_{ni} KL((Ax)_i, (Az)_i). \tag{26.59}$$

Therefore the distance $G_n(x, z)$ is minimized, as a function of z, by $z = x$. Now we minimize $G_n(x, z)$ as a function of x. The following lemma shows that the answer is

$$x_j = z'_j = z_j \exp\Big(\gamma_j \delta_n \sum_{i \in B_n} \alpha_{ni} A_{ij} \log \frac{b_i}{(Az)_i}\Big). \tag{26.60}$$

Lemma 26.6 *For each x and z we have*

$$G_n(x, z) = G_n(z', z) + \sum_{j=1}^{J} \gamma_j^{-1} KL(x_j, z'_j). \tag{26.61}$$

It is clear that $(x^k)' = x^{k+1}$ for all k.

Now let $b = Au$ for some nonnegative vector u. We calculate $G_n(u, x^k)$ in two ways: using the definition we have

$$G_n(u, x^k) = \sum_{j=1}^{J} \gamma_j^{-1} KL(u_j, x_j^k) - \delta_n \sum_{i \in B_n} \alpha_{ni} KL(b_i, (Ax^k)_i), \tag{26.62}$$

while using Lemma 26.61 we find that

$$G_n(u, x^k) = G_n(x^{k+1}, x^k) + \sum_{j=1}^{J} \gamma_j^{-1} KL(u_j, x_j^{k+1}). \tag{26.63}$$

Therefore

$$\sum_{j=1}^{J} \gamma_j^{-1} KL(u_j, x_j^k) - \sum_{j=1}^{J} \gamma_j^{-1} KL(u_j, x_j^{k+1}) \tag{26.64}$$

$$= G_n(x^{k+1}, x^k) + \delta_n \sum_{i \in B_n} \alpha_{ni} KL(b_i, (Ax^k)_i). \tag{26.65}$$

We conclude several things from this.

First, the sequence $\{\sum_{j=1}^{J} \gamma_j^{-1} KL(u_j, x_j^k)\}$ is decreasing, so that the sequences $\{G_n(x^{k+1}, x^k)\}$ and $\{\delta_n \sum_{i \in B_n} \alpha_{ni} KL(b_i, (Ax^k)_i)\}$ converge to zero. Therefore the sequence $\{x^k\}$ is bounded and we may select an arbitrary cluster point x^*. It follows that $b = Ax^*$. We may therefore replace the generic solution u with x^* to find that $\{\sum_{j=1}^{J} \gamma_j^{-1} KL(x_j^*, x_j^k)\}$ is a decreasing sequence; but since a subsequence converges to zero, the entire sequence must converge to zero. Therefore $\{x^k\}$ converges to the solution x^*.

Finally, since the right side of Equation (26.65) does not depend on the particular choice of solution we made, neither does the left side. By *telescoping* we conclude that

$$\sum_{j=1}^{J} \gamma_j^{-1} KL(u_j, x_j^0) - \sum_{j=1}^{J} \gamma_j^{-1} KL(u_j, x_j^*) \tag{26.66}$$

is also independent of the choice of u. Consequently, minimizing the function $\sum_{j=1}^{J} \gamma_j^{-1} KL(u_j, x_j^0)$ over all solutions u is equivalent to minimizing $\sum_{j=1}^{J} \gamma_j^{-1} KL(u_j, x_j^*)$ over all solutions u; but the solution to the latter problem is obviously $u = x^*$. This completes the proof.

Chapter 27

Appendix: Eigenvalue Bounds

27.1 Chapter Summary

The ART is a sequential algorithm, using only a single equation from the system $Ax = b$ at each step of the iteration. In this chapter we consider iterative procedures for solving $Ax = b$ in which several or all of the equations are used at each step. Such methods are called *block-iterative* and *simultaneous* algorithms, respectively. We survey a number of these block-iterative methods. We obtain upper bounds on the spectral radius of

positive-definite matrices and use these bounds in the selection of parameters in the iterative methods.

27.2 Introduction and Notation

We are concerned here with iterative methods for solving, at least approximately, the system of I linear equations in J unknowns symbolized by $Ax = b$. In the applications of interest to us, such as medical imaging, both I and J are quite large, making the use of iterative methods the only feasible approach. It is also typical of such applications that the matrix A is sparse, that is, has relatively few non-zero entries. Therefore, iterative methods that exploit this sparseness to accelerate convergence are of special interest to us.

The *algebraic reconstruction technique* (ART) of Gordon, et al. [151] is a *sequential* method; at each step only one equation is used. The current vector x^{k-1} is projected orthogonally onto the hyperplane corresponding to that single equation, to obtain the next iterate x^k. The iterative step of the ART is

$$x_j^k = x_j^{k-1} + \overline{A_{ij}}\left(\frac{b_i - (Ax^{k-1})_i}{\sum_{t=1}^J |A_{it}|^2}\right), \tag{27.1}$$

where $i = k(\mathrm{mod}\, I)$. The sequence $\{x^k\}$ converges to the solution closest to x^0 in the consistent case, but only converges subsequentially to a limit cycle in the inconsistent case.

Cimmino's method [94] is a *simultaneous* method, in which all the equations are used at each step. The current vector x^{k-1} is projected orthogonally onto each of the hyperplanes and these projections are averaged to obtain the next iterate x^k. The iterative step of Cimmino's method is

$$x_j^k = \frac{1}{I}\sum_{i=1}^I \left(x_j^{k-1} + \overline{A_{ij}}\left(\frac{b_i - (Ax^{k-1})_i}{\sum_{t=1}^J |A_{it}|^2}\right)\right),$$

which can also be written as

$$x_j^k = x_j^{k-1} + \sum_{i=1}^I \overline{A_{ij}}\left(\frac{b_i - (Ax^{k-1})_i}{I\sum_{t=1}^J |A_{it}|^2}\right). \tag{27.2}$$

Landweber's iterative scheme [187] with

$$x^k = x^{k-1} + B^\dagger(d - Bx^{k-1}), \tag{27.3}$$

converges to the least-squares solution of $Bx = d$ closest to x^0, provided that the largest singular value of B does not exceed one. If we let B be the matrix with entries

$$B_{ij} = A_{ij} / \sqrt{I \sum_{t=1}^{J} |A_{it}|^2},$$

and define

$$d_i = b_i / \sqrt{I \sum_{t=1}^{J} |A_{it}|^2},$$

then, since the trace of the matrix $BB^\dagger$ is one, convergence of Cimmino's method follows. However, using the trace in this way to estimate the largest singular value of a matrix usually results in an estimate that is far too large, particularly when A is large and sparse, and therefore in an iterative algorithm with unnecessarily small step sizes.

The appearance of the term

$$I \sum_{t=1}^{J} |A_{it}|^2$$

in the denominator of Cimmino's method suggested to Censor et al. [85] that, when A is sparse, this denominator might be replaced with

$$\sum_{t=1}^{J} s_t |A_{it}|^2,$$

where s_t denotes the number of non-zero entries in the tth column of A. The resulting iterative method is the *component-averaging* (CAV) iteration. Convergence of the CAV method was established by showing that no singular value of the matrix B exceeds one, where B has the entries

$$B_{ij} = A_{ij} / \sqrt{\sum_{t=1}^{J} s_t |A_{it}|^2}.$$

In [69] we extended this result, to show that no eigenvalue of $A^\dagger A$ exceeds the maximum of the numbers

$$p_i = \sum_{t=1}^{J} s_t |A_{it}|^2.$$

Convergence of CAV then follows, as does convergence of several other methods, including the ART, Landweber's method, the SART [5], the block-iterative CAV (BICAV) [86], the CARP1 method of Gordon and

Gordon [152], a block-iterative variant of CARP1 obtained from the DROP method of Censor et al. [83], and the SIRT method [257].

For a positive integer N with $1 \leq N \leq I$, we let $B_1, ..., B_N$ be not necessarily disjoint subsets of the set $\{i = 1, ..., I\}$; the subsets B_n are called *blocks*. We then let A_n be the matrix and b^n the vector obtained from A and b, respectively, by removing all the rows except for those whose index i is in the set B_n. For each n, we let s_{nt} be the number of non-zero entries in the tth column of the matrix A_n, s_n the maximum of the s_{nt}, s the maximum of the s_t, and $L_n = \rho(A_n^\dagger A_n)$ be the spectral radius, or largest eigenvalue, of the matrix $A_n^\dagger A_n$, with $L = \rho(A^\dagger A)$. We denote by A_i the ith row of the matrix A, and by ν_i the length of A_i, so that

$$\nu_i^2 = \sum_{j=1}^{J} |A_{ij}|^2.$$

27.3 Cimmino's Algorithm

The ART seeks a solution of $Ax = b$ by projecting the current vector x^{k-1} orthogonally onto the next hyperplane $H(a^{i(k)}, b_{i(k)})$ to get x^k; here $i(k) = k(\mathrm{mod}\,)I$. In Cimmino's algorithm, we project the current vector x^{k-1} onto each of the hyperplanes and then average the result to get x^k. The algorithm begins at $k = 1$, with an arbitrary x^0; the iterative step is then

$$x^k = \frac{1}{I} \sum_{i=1}^{I} P_i x^{k-1}, \tag{27.4}$$

where P_i is the orthogonal projection onto $H(a^i, b_i)$. The iterative step can then be written as

$$x_j^k = x_j^{k-1} + \frac{1}{I} \sum_{i=1}^{I} \left(\frac{\overline{A_{ij}}(b_i - (Ax^{k-1})_i)}{\nu_i^2} \right). \tag{27.5}$$

As we saw in our discussion of the ART, when the system $Ax = b$ has no solutions, the ART does not converge to a single vector, but to a limit cycle. One advantage of many simultaneous algorithms, such as Cimmino's, is that they do converge to the least squares solution in the inconsistent case.

When $\nu_i = 1$ for all i, Cimmino's algorithm has the form $x^{k+1} = Tx^k$, for the operator T given by

$$Tx = (I - \frac{1}{I} A^\dagger A)x + \frac{1}{I} A^\dagger b.$$

Experience with Cimmino's algorithm shows that it is slow to converge. In the next section we consider how we might accelerate the algorithm.

27.4 The Landweber Algorithms

For simplicity, we assume, in this section, that $\nu_i = 1$ for all i. The Landweber algorithm [187, 18], with the iterative step

$$x^k = x^{k-1} + \gamma A^\dagger(b - Ax^{k-1}), \tag{27.6}$$

converges to the least squares solution closest to the starting vector x^0, provided that $0 < \gamma < 2/\lambda_{max}$, where λ_{max} is the largest eigenvalue of the nonnegative-definite matrix $A^\dagger A$. Loosely speaking, the larger γ is, the faster the convergence. However, precisely because A is large, calculating the matrix $A^\dagger A$, not to mention finding its largest eigenvalue, can be prohibitively expensive. The matrix A is said to be sparse if most of its entries are zero. Useful upper bounds for λ_{max} are then given by Theorems 27.1 and 27.6.

27.4.1 Finding the Optimum γ

The operator

$$Tx = x + \gamma A^\dagger(b - Ax) = (I - \gamma A^\dagger A)x + \gamma A^\dagger b$$

is affine linear and is av if and only if its linear part, the Hermitian matrix

$$B = I - \gamma A^\dagger A,$$

is av. To guarantee this we need $0 \leq \gamma < 2/\lambda_{max}$. Should we always try to take γ near its upper bound, or is there an optimum value of γ? To answer this question we consider the eigenvalues of B for various values of γ.

Lemma 27.1 *If $\gamma < 0$, then none of the eigenvalues of B is less than one.*

Lemma 27.2 *For*

$$0 \leq \gamma \leq \frac{2}{\lambda_{max} + \lambda_{min}}, \tag{27.7}$$

we have

$$\rho(B) = 1 - \gamma\lambda_{min}; \tag{27.8}$$

the smallest value of $\rho(B)$ *occurs when*

$$\gamma = \frac{2}{\lambda_{max} + \lambda_{min}}, \tag{27.9}$$

and equals

$$\frac{\lambda_{max} - \lambda_{min}}{\lambda_{max} + \lambda_{min}}. \tag{27.10}$$

Similarly, for

$$\gamma \geq \frac{2}{\lambda_{max} + \lambda_{min}}, \tag{27.11}$$

we have

$$\rho(B) = \gamma\lambda_{max} - 1; \tag{27.12}$$

the smallest value of $\rho(B)$ *occurs when*

$$\gamma = \frac{2}{\lambda_{max} + \lambda_{min}}, \tag{27.13}$$

and equals

$$\frac{\lambda_{max} - \lambda_{min}}{\lambda_{max} + \lambda_{min}}. \tag{27.14}$$

We see from this lemma that, if $0 \leq \gamma < 2/\lambda_{max}$, and $\lambda_{min} > 0$, then $\|B\|_2 = \rho(B) < 1$, so that B is sc. We minimize $\|B\|_2$ by taking

$$\gamma = \frac{2}{\lambda_{max} + \lambda_{min}}, \tag{27.15}$$

in which case we have

$$\|B\|_2 = \frac{\lambda_{max} - \lambda_{min}}{\lambda_{max} + \lambda_{min}} = \frac{c-1}{c+1}, \tag{27.16}$$

for $c = \lambda_{max}/\lambda_{min}$, the *condition number* of the positive-definite matrix $A^\dagger A$. The closer c is to one, the smaller the norm $\|B\|_2$, and the faster the convergence.

On the other hand, if $\lambda_{min} = 0$, then $\rho(B) = 1$ for all γ in the interval $(0, 2/\lambda_{max})$. The matrix B is still av, but it is no longer sc. For example, consider the orthogonal projection P_0 onto the hyperplane $H_0 = H(a, 0)$, where $\|a\|_2 = 1$. This operator can be written

$$P_0 = I - aa^\dagger. \tag{27.17}$$

The largest eigenvalue of $aa^\dagger$ is $\lambda_{max} = 1$; the remaining ones are zero. The relaxed projection operator

$$B = I - \gamma aa^\dagger \tag{27.18}$$

has $\rho(B) = 1 - \gamma > 1$, if $\gamma < 0$, and for $\gamma \geq 0$, we have $\rho(B) = 1$. The operator B is av, in fact, it is fne, but it is not sc.

27.4.2 The Projected Landweber Algorithm

When we require a nonnegative approximate solution x for the real system $Ax = b$ we can use a modified version of the Landweber algorithm, called the projected Landweber algorithm [18], in this case having the iterative step

$$x^{k+1} = (x^k + \gamma A^\dagger(b - Ax^k))_+, \tag{27.19}$$

where, for any real vector a, we denote by $(a)_+$ the nonnegative vector whose entries are those of a, for those that are nonnegative, and are zero otherwise. The projected Landweber algorithm converges to a vector that minimizes $\|Ax - b\|_2$ over all nonnegative vectors x, for the same values of γ.

The projected Landweber algorithm is actually more general. For any closed, nonempty convex set C in X, define the iterative sequence

$$x^{k+1} = P_C(x^k + \gamma A^\dagger(b - Ax^k)). \tag{27.20}$$

This sequence converges to a minimizer of the function $\|Ax - b\|_2$ over all x in C, whenever such minimizers exist.

Both the Landweber and projected Landweber algorithms are special cases of the CQ algorithm [62], which, in turn, is a special case of the more general iterative fixed point algorithm, the Krasnoselskii/Mann (KM) method, with convergence governed by the KM Theorem (see [70]).

27.5 Some Upper Bounds for L

For the iterative algorithms we shall consider here, having a good upper bound for the largest eigenvalue of the matrix $A^\dagger A$ is important. In the applications of interest, principally medical image processing, the matrix A is large; even calculating $A^\dagger A$, not to mention computing eigenvalues, is prohibitively expensive. In addition, the matrix A is typically sparse, but $A^\dagger A$ will not be, in general. In this section we present upper bounds for L that are particularly useful when A is sparse and do not require the calculation of $A^\dagger A$.

27.5.1 Earlier Work

Many of the concepts we study in computational linear algebra were added to the mathematical toolbox relatively recently, as this area blossomed with the growth of electronic computers. Based on my brief investigations into the history of matrix theory, I believe that the concept of

a norm of a matrix was not widely used prior to about 1945. This was recently confirmed when I read the paper [153]; as pointed out there, the use of matrix norms became an important part of numerical linear algebra only after the publication of [260]. Prior to the late 1940's a number of papers were published that established upper bounds on $\rho(A)$, for general square matrix A. As we now can see, several of these results are immediate consequences of the fact that $\rho(A) \leq \|A\|$, for any induced matrix norm. We give two examples.

For a given N by N matrix A, let

$$C_n = \sum_{m=1}^{N} |A_{mn}|,$$

$$R_m = \sum_{n=1}^{N} |A_{mn}|,$$

and C and R the maxima of C_n and R_m, respectively. We now know that $C = \|A\|_1$, and $R = \|A\|_\infty$, but the earlier authors did not.

In 1930 Browne [32] proved the following theorem.

Theorem 27.1 (Browne) *Let λ be any eigenvalue of A. Then*

$$|\lambda| \leq \frac{1}{2}(C + R).$$

In 1944 Farnell [130] published the following theorems.

Theorem 27.2 (Farnell I) *For any eigenvalue λ of A we have*

$$|\lambda| \leq \sqrt{CR}.$$

Theorem 27.3 (Farnell II) *Let*

$$r_m = \sum_{n=1}^{N} |A_{mn}|^2,$$

and

$$c_m = \sum_{n=1}^{N} |A_{nm}|^2.$$

Then, for any eigenvalue λ of A, we have

$$|\lambda| \leq \sqrt{\sum_{m=1}^{N} \sqrt{r_m c_m}}.$$

In 1946 Brauer [28] proved the following theorem.

Theorem 27.4 (Brauer) *For any eigenvalue λ of A, we have*

$$|\lambda| \leq \min\{C, R\}.$$

Ex. 27.1 *Prove Theorems 27.1, 27.2, and 27.4 using properties of matrix norms. Can you also prove Theorem 27.3 this way?*

Let A be an arbitrary rectangular complex matrix. Since the largest singular value of A is the square root of the maximum eigenvalue of the square matrix $S = A^\dagger A$, we could use the inequality

$$\rho(A^\dagger A) = \|A^\dagger A\|_2 \leq \|A^\dagger A\|,$$

for any induced matrix norm, to establish an upper bound for the singular values of A. However, that bound would be in terms of the entries of $A^\dagger A$, not of A itself. In what follows we obtain upper bounds on the singular values of A in terms of the entries of A itself.

Ex. 27.2 *Let A be an arbitrary rectangular matrix. Prove that no singular value of A exceeds $\sqrt{\|A\|_1 \|A\|_\infty}$.*

We see from this exercise that Farnell (I) does generalize to arbitrary rectangular matrices and singular values. Brauer's Theorem 27.4 may suggest that no singular value of a rectangular matrix A exceeds the minimum of $\|A\|_1$ and $\|A\|_\infty$, but this is not true. Consider the matrix A whose SVD is given by

$$A = \begin{bmatrix} 4 & 3 \\ 8 & 6 \\ 8 & 6 \end{bmatrix} = \begin{bmatrix} 1/3 & 2/3 & 2/3 \\ 2/3 & -2/3 & 1/3 \\ 2/3 & 1/3 & -2/3 \end{bmatrix} \begin{bmatrix} 15 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 4/5 & 3/5 \\ 3/5 & -4/5 \end{bmatrix}.$$

The largest singular value of A is 15, $\|A\|_1 = 20$, $\|A\|_\infty = 14$, and we do have

$$15 \leq \sqrt{(20)(14)},$$

but we do not have

$$15 \leq \min\{20, 14\} = 14.$$

27.5.2 Our Basic Eigenvalue Inequality

In [257] van der Sluis and van der Vorst show that certain rescaling of the matrix A results in none of the eigenvalues of $A^\dagger A$ exceeding one. A modification of their proof leads to upper bounds on the eigenvalues of the original $A^\dagger A$ ([69]). For any a in the interval $[0, 2]$ let

$$c_{aj} = c_{aj}(A) = \sum_{i=1}^{I} |A_{ij}|^a,$$

$$r_{ai} = r_{ai}(A) = \sum_{j=1}^{J} |A_{ij}|^{2-a},$$

and c_a and r_a the maxima of the c_{aj} and r_{ai}, respectively. We prove the following theorem.

Theorem 27.5 *For any a in the interval $[0, 2]$, no eigenvalue of the matrix $A^\dagger A$ exceeds the maximum of*

$$\sum_{j=1}^{J} c_{aj}|A_{ij}|^{2-a},$$

over all i, nor the maximum of

$$\sum_{i=1}^{I} r_{ai}|A_{ij}|^{a},$$

over all j. Therefore, no eigenvalue of $A^\dagger A$ exceeds $c_a r_a$.

Proof: Let $A^\dagger Av = \lambda v$, and let $w = Av$. Then we have

$$\|A^\dagger w\|_2^2 = \lambda \|w\|_2^2.$$

Applying Cauchy's Inequality, we obtain

$$\Big|\sum_{i=1}^{I} \overline{A_{ij}} w_i\Big|^2 \leq \Big(\sum_{i=1}^{I} |A_{ij}|^{a/2}|A_{ij}|^{1-a/2}|w_i|\Big)^2$$

$$\leq \Big(\sum_{i=1}^{I} |A_{ij}|^{a}\Big)\Big(\sum_{i=1}^{I} |A_{ij}|^{2-a}|w_i|^2\Big).$$

Therefore,

$$\|A^\dagger w\|_2^2 \leq \sum_{j=1}^{J} \Big(c_{aj}(\sum_{i=1}^{I} |A_{ij}|^{2-a}|w_i|^2)\Big) = \sum_{i=1}^{I} \Big(\sum_{j=1}^{J} c_{aj}|A_{ij}|^{2-a}\Big)|w_i|^2$$

$$\leq \max_i \Big(\sum_{j=1}^{J} c_{aj}|A_{ij}|^{2-a}\Big)\|w\|^2.$$

The remaining two assertions follow in similar fashion. ∎

As a corollary, we obtain the following eigenvalue inequality, which is central to our discussion.

Corollary 27.1 *For each $i = 1, 2, ..., I$, let*

$$p_i = \sum_{j=1}^{J} s_j |A_{ij}|^2,$$

and let p be the maximum of the p_i. Then $L \leq p$.

Proof: Take $a = 0$. Then, using the convention that $0^0 = 0$, we have $c_{0j} = s_j$. ∎

Corollary 27.2 *([62]; [256], Th. 4.2) If $\sum_{j=1}^{J} |A_{ij}|^2 \leq 1$ for each i, then $L \leq s$.*

Proof: For all i we have

$$p_i = \sum_{j=1}^{J} s_j |A_{ij}|^2 \leq s \sum_{j=1}^{J} |A_{ij}|^2 \leq s.$$

Therefore,

$$L \leq p \leq s.$$

∎

The next corollary gives Inequality (6.38) that we saw earlier.

Corollary 27.3 *Selecting $a = 1$, we have*

$$L = \|A\|_2^2 \leq \|A\|_1 \|A\|_\infty = c_1 r_1.$$

Therefore, the largest singular value of A does not exceed $\sqrt{\|A\|_1 \|A\|_\infty}$.

Corollary 27.4 *Selecting $a = 2$, we have*

$$L = \|A\|_2^2 \leq \|A\|_F^2,$$

where $\|A\|_F$ denotes the Frobenius norm of A.

Corollary 27.5 *Let G be the matrix with entries*

$$G_{ij} = A_{ij} \sqrt{\alpha_i} \sqrt{\beta_j},$$

where

$$\alpha_i \leq \Big(\sum_{j=1}^{J} s_j \beta_j |A_{ij}|^2 \Big)^{-1},$$

for all i. Then $\rho(G^\dagger G) \leq 1$.

Proof: We have

$$\sum_{j=1}^{J} s_j |G_{ij}|^2 = \alpha_i \sum_{j=1}^{J} s_j \beta_j |A_{ij}|^2 \leq 1,$$

for all i. The result follows from Corollary 27.1. ∎

Corollary 27.6 *If $\sum_{j=1}^{J} s_j |A_{ij}|^2 \leq 1$ for all i, then $L \leq 1$.*

Corollary 27.7 *If $0 < \gamma_i \leq p_i^{-1}$ for all i, then the matrix B with entries $B_{ij} = \sqrt{\gamma_i} A_{ij}$ has $\rho(B^\dagger B) \leq 1$.*

Proof: We have

$$\sum_{j=1}^{J} s_j |B_{ij}|^2 = \gamma_i \sum_{j=1}^{J} s_j |A_{ij}|^2 = \gamma_i p_i \leq 1.$$

Therefore, $\rho(B^\dagger B) \leq 1$, according to the theorem. ∎

Corollary 27.8 *If, for some a in the interval $[0, 2]$, we have*

$$\alpha_i \leq r_{ai}^{-1}, \tag{27.21}$$

for each i, and

$$\beta_j \leq c_{aj}^{-1}, \tag{27.22}$$

for each j, then, for the matrix G with entries

$$G_{ij} = A_{ij} \sqrt{\alpha_i} \sqrt{\beta_j},$$

no eigenvalue of $G^\dagger G$ exceeds one.

Proof: We calculate $c_{aj}(G)$ and $r_{ai}(G)$ and find that

$$c_{aj}(G) \leq \Big(\max_i \alpha_i^{a/2} \Big) \beta_j^{a/2} \sum_{i=1}^{I} |A_{ij}|^a = \Big(\max_i \alpha_i^{a/2} \Big) \beta_j^{a/2} c_{aj}(A),$$

and

$$r_{ai}(G) \leq \Big(\max_j \beta_j^{1-a/2} \Big) \alpha_i^{1-a/2} r_{ai}(A).$$

Therefore, applying the inequalities (27.21) and (27.22), we have

$$c_{aj}(G) r_{ai}(G) \leq 1,$$

for all i and j. Consequently, $\rho(G^\dagger G) \leq 1$. ∎

27.5.3 Another Upper Bound for L

The next theorem ([62]) provides another upper bound for L that is useful when A is sparse. As previously, for each i and j, we let $e_{ij} = 1$, if A_{ij} is not zero, and $e_{ij} = 0$, if $A_{ij} = 0$. Let $0 < \nu_i = \sqrt{\sum_{j=1}^{J} |A_{ij}|^2}$, $\sigma_j = \sum_{i=1}^{I} e_{ij}\nu_i^2$, and σ be the maximum of the σ_j.

Theorem 27.6 *([62]) No eigenvalue of $A^\dagger A$ exceeds σ.*

Proof: Let $A^\dagger Av = cv$, for some non-zero vector v and scalar c. With $w = Av$, we have

$$w^\dagger AA^\dagger w = cw^\dagger w.$$

Then

$$\Big|\sum_{i=1}^{I} \overline{A_{ij}} w_i\Big|^2 = \Big|\sum_{i=1}^{I} \overline{A_{ij}} e_{ij}\nu_i \frac{w_i}{\nu_i}\Big|^2 \le \Big(\sum_{i=1}^{I} |A_{ij}|^2 \frac{|w_i|^2}{\nu_i^2}\Big)\Big(\sum_{i=1}^{I} \nu_i^2 e_{ij}\Big)$$

$$= \Big(\sum_{i=1}^{I} |A_{ij}|^2 \frac{|w_i|^2}{\nu_i^2}\Big)\sigma_j \le \sigma\Big(\sum_{i=1}^{I} |A_{ij}|^2 \frac{|w_i|^2}{\nu_i^2}\Big).$$

Therefore, we have

$$cw^\dagger w = w^\dagger AA^\dagger w = \sum_{j=1}^{J} \Big|\sum_{i=1}^{I} \overline{A_{ij}} w_i\Big|^2$$

$$\le \sigma \sum_{j=1}^{J} \Big(\sum_{i=1}^{I} |A_{ij}|^2 \frac{|w_i|^2}{\nu_i^2}\Big) = \sigma \sum_{i=1}^{I} |w_i|^2 = \sigma w^\dagger w.$$

We conclude that $c \le \sigma$. ∎

Corollary 27.9 *Let the rows of A have Euclidean length one. Then no eigenvalue of $A^\dagger A$ exceeds the maximum number of non-zero entries in any column of A.*

Proof: We have $\nu_i^2 = \sum_{j=1}^{J} |A_{ij}|^2 = 1$, for each i, so that $\sigma_j = s_j$ is the number of non-zero entries in the jth column of A, and $\sigma = s$ is the maximum of the σ_j. ∎

Corollary 27.10 *Let ν be the maximum Euclidean length of any row of A and s the maximum number of non-zero entries in any column of A. Then $L \le \nu^2 s$.*

When the rows of A have length one, it is easy to see that $L \le I$, so the choice of $\gamma = \frac{1}{I}$ in the Landweber algorithm, which gives Cimmino's algorithm [94], is acceptable, although perhaps much too small.

The proof of Theorem 27.6 is based on results presented by Arnold Lent in informal discussions with Gabor Herman, Yair Censor, Rob Lewitt and me at MIPG in Philadelphia in the late 1990's.

27.6 Eigenvalues and Norms: A Summary

It is helpful, at this point, to summarize the main facts concerning eigenvalues and norms. Throughout this section A will denote an arbitrary matrix, S an arbitrary square matrix, and H an arbitrary Hermitian matrix. We denote by $\|A\|$ an arbitrary induced matrix norm of A.

Here are some of the things we now know:

- 1. $\rho(S^2) = \rho(S)^2$;
- 2. $\rho(S) \leq \|S\|$, for any matrix norm;
- 3. $\rho(H) = \|H\|_2 \leq \|H\|$, for any matrix norm;
- 4. $\|A\|_2^2 = \rho(A^\dagger A) = \|A^\dagger A\|_2 \leq \|A^\dagger A\|$;
- 5. $\|A^\dagger A\|_1 \leq \|A^\dagger\|_1 \|A\|_1 = \|A\|_\infty \|A\|_1$;
- 6. $\|A\|_2^2 \leq \|A\|_1 \|A\|_\infty$;
- 7. $\rho(S) \leq \min\{\|S\|_1, \|S\|_\infty\}$;
- 8. if $\sum_{j=1}^J |A_{ij}|^2 \leq 1$, for all i, then $\|A\|_2^2 \leq s$, where s is the largest number of non-zero entries in any column of A.

27.7 The Basic Convergence Theorem

The following theorem is a basic convergence result concerning block-iterative ART algorithms.

Theorem 27.7 *Let $L_n \leq 1$, for $n = 1, 2, ..., N$. If the system $Ax = b$ is consistent, then, for any starting vector x^0, and with $n = n(k) = k(\mathrm{mod}\, N)$ and $\lambda_k \in [\epsilon, 2 - \epsilon]$ for all k, the sequence $\{x^k\}$ with iterative step*

$$x^k = x^{k-1} + \lambda_k A_n^\dagger (b^n - A_n x^{k-1}) \tag{27.23}$$

converges to the solution of $Ax = b$ for which $\|x - x^0\|_2$ is minimized.

We begin with the following lemma.

Lemma 27.3 *Let T be any (not necessarily linear) operator on $\mathbb{R}^J$, and $S = I - T$, where I denotes the identity operator. Then, for any x and y, we have*

$$\|x - y\|_2^2 - \|Tx - Ty\|_2^2 = 2\langle Sx - Sy, x - y\rangle - \|Sx - Sy\|^2. \quad (27.24)$$

The proof is a simple calculation and we omit it here.

Proof of Theorem 27.7: Let $Az = b$. Applying Equation (27.24) to the operator

$$Tx = x + \lambda_k A_n^\dagger (b^n - A_n x),$$

we obtain

$$\|z - x^{k-1}\|_2^2 - \|z - x^k\|_2^2 = 2\lambda_k \|b^n - A_n x^{k-1}\|_2^2 - \lambda_k^2 \|A_n^\dagger b^n - A_n^\dagger A_n x^{k-1}\|_2^2. \quad (27.25)$$

Since $L_n \leq 1$, it follows that

$$\|A_n^\dagger b^n - A_n^\dagger A_n x^{k-1}\|_2^2 \leq \|b^n - A_n x^{k-1}\|_2^2.$$

Therefore,

$$\|z - x^{k-1}\|_2^2 - \|z - x^k\|_2^2 \geq (2\lambda_k - \lambda_k^2)\|b^n - A_n x^{k-1}\|_2^2,$$

from which we draw several conclusions:

- the sequence $\{\|z - x^k\|_2\}$ is decreasing;
- the sequence $\{\|b^n - A_n x^{k-1}\|_2\}$ converges to zero.

In addition, for fixed $n = 1, ..., N$ and $m \to \infty$,

- the sequence $\{\|b^n - A_n x^{mN+n-1}\|_2\}$ converges to zero;
- the sequence $\{x^{mN+n}\}$ is bounded.

Let $x^{*,1}$ be a cluster point of the sequence $\{x^{mN+1}\}$; then there is subsequence $\{x^{m_r N+1}\}$ converging to $x^{*,1}$. The sequence $\{x^{m_r N+2}\}$ is also bounded, and we select a cluster point $x^{*,2}$. Continuing in this fashion, we obtain cluster points $x^{*,n}$, for $n = 1, ..., N$. From the conclusions reached previously, we can show that $x^{*,n} = x^{*,n+1} = x^*$, for $n = 1, 2, ..., N-1$, and $Ax^* = b$. Replacing the generic solution $\hat{x}$ with the solution x^*, we see that the sequence $\{\|x^* - x^k\|_2\}$ is decreasing. But, subsequences of this sequence converge to zero, so the entire sequence converges to zero, and so $x^k \to x^*$.

Now we show that x^* is the solution of $Ax = b$ that minimizes $\|x - x^0\|_2$. Since $x^k - x^{k-1}$ is in the range of $A^\dagger$ for all k, so is $x^* - x^0$, from which it follows that x^* is the solution minimizing $\|x - x^0\|_2$. Another way to

get this result is to use Equation (27.25). Since the right side of Equation (27.25) is independent of the choice of solution, so is the left side. Summing both sides over the index k reveals that the difference

$$\|x - x^0\|_2^2 - \|x - x^*\|_2^2$$

is independent of the choice of solution. Consequently, minimizing $\|x-x^0\|_2$ over all solutions x is equivalent to minimizing $\|x - x^*\|_2$ over all solutions x; the solution to the latter problem is clearly $x = x^*$. ∎

27.8 Simultaneous Iterative Algorithms

In this section we apply the previous theorems to obtain convergence of several simultaneous iterative algorithms for linear systems.

27.8.1 The General Simultaneous Iterative Scheme

In this section we are concerned with simultaneous iterative algorithms having the following iterative step:

$$x_j^k = x_j^{k-1} + \lambda_k \sum_{i=1}^{I} \gamma_{ij} \overline{A_{ij}} (b_i - (Ax^{k-1})_i), \tag{27.26}$$

with $\lambda_k \in [\epsilon, 1]$ and the choices of the parameters γ_{ij} that guarantee convergence. Although we cannot prove convergence for this most general iterative scheme, we are able to prove the following theorems for the separable case of $\gamma_{ij} = \alpha_i \beta_j$.

Theorem 27.8 *If, for some a in the interval $[0, 2]$, we have*

$$\alpha_i \leq r_{ai}^{-1}, \tag{27.27}$$

for each i, and

$$\beta_j \leq c_{aj}^{-1}, \tag{27.28}$$

for each j, then the sequence $\{x^k\}$ given by Equation (27.26) converges to the minimizer of the proximity function

$$\sum_{i=1}^{I} \alpha_i |b_i - (Ax)_i|^2$$

for which

$$\sum_{j=1}^{J} \beta_j^{-1} |x_j - x_j^0|^2$$

is minimized.

Proof: For each i and j, let

$$G_{ij} = \sqrt{\alpha_i}\sqrt{\beta_j} A_{ij},$$

$$z_j = x_j / \sqrt{\beta_j},$$

and

$$d_i = \sqrt{\alpha_i} b_i.$$

Then $Ax = b$ if and only if $Gz = d$. From Corollary 27.8 we have that $\rho(G^\dagger G) \leq 1$. Convergence then follows from Theorem 27.7. ∎

Corollary 27.11 *Let $\gamma_{ij} = \alpha_i \beta_j$, for positive α_i and β_j. If*

$$\alpha_i \leq \Big(\sum_{j=1}^{J} s_j \beta_j |A_{ij}|^2 \Big)^{-1}, \tag{27.29}$$

for each i, then the sequence $\{x^k\}$ in (27.26) converges to the minimizer of the proximity function

$$\sum_{i=1}^{I} \alpha_i |b_i - (Ax)_i|^2$$

for which

$$\sum_{j=1}^{J} \beta_j^{-1} |x_j - x_j^0|^2$$

is minimized.

Proof: We know from Corollary 27.5 that $\rho(G^\dagger G) \leq 1$. ∎

We now obtain convergence for several known algorithms as corollaries to the previous theorems.

27.8.2 The SIRT Algorithm

Corollary 27.12 *([257]) For some a in the interval $[0, 2]$ let $\alpha_i = r_{ai}^{-1}$ and $\beta_j = c_{aj}^{-1}$. Then the sequence $\{x^k\}$ in (27.26) converges to the minimizer of the proximity function*

$$\sum_{i=1}^{I} \alpha_i |b_i - (Ax)_i|^2$$

for which

$$\sum_{j=1}^{J} \beta_j^{-1} |x_j - x_j^0|^2$$

is minimized.

For the case of $a = 1$, the iterative step becomes

$$x_j^k = x_j^{k-1} + \sum_{i=1}^{I} \left(\frac{\overline{A_{ij}}(b_i - (Ax^{k-1})_i)}{(\sum_{t=1}^{J} |A_{it}|)(\sum_{m=1}^{I} |A_{mj}|)} \right),$$

which was considered in [159]. The SART algorithm [5] is a special case, in which it is assumed that $A_{ij} \geq 0$, for all i and j.

27.8.3 The CAV Algorithm

Corollary 27.13 *If $\beta_j = 1$ and α_i satisfies*

$$0 < \alpha_i \leq \Big(\sum_{j=1}^{J} s_j |A_{ij}|^2 \Big)^{-1},$$

for each i, then the algorithm with the iterative step

$$x^k = x^{k-1} + \lambda_k \sum_{i=1}^{I} \alpha_i (b_i - (Ax^{k-1})_i) A_i^\dagger \tag{27.30}$$

converges to the minimizer of

$$\sum_{i=1}^{I} \alpha_i |b_i - (Ax^{k-1})_i|^2$$

for which $\|x - x^0\|$ is minimized.

When

$$\alpha_i = \Big(\sum_{j=1}^{J} s_j |A_{ij}|^2 \Big)^{-1},$$

for each i, this is the relaxed *component-averaging* (CAV) method of Censor et al. [85].

27.8.4 The Landweber Algorithm

When $\beta_j = 1$ and $\alpha_i = \alpha$ for all i and j, we have the relaxed Landweber algorithm. The convergence condition in Equation (27.21) becomes

$$\alpha \leq \Big(\sum_{j=1}^{J} s_j |A_{ij}|^2 \Big)^{-1} = p_i^{-1}$$

for all i, so $\alpha \leq p^{-1}$ suffices for convergence. Actually, the sequence $\{x^k\}$ converges to the minimizer of $\|Ax - b\|_2$ for which the distance $\|x - x^0\|_2$ is minimized, for any starting vector x^0, when $0 < \alpha < 1/L$. Easily obtained estimates of L are usually over-estimates, resulting in overly conservative choices of α. For example, if A is first normalized so that $\sum_{j=1}^{J} |A_{ij}|^2 = 1$ for each i, then the trace of $A^\dagger A$ equals I, which tells us that $L \leq I$. But this estimate, which is the one used in Cimmino's method [94], is far too large when A is sparse.

27.8.5 The Simultaneous DROP Algorithm

Corollary 27.14 *Let* $0 < w_i \leq 1$,

$$\alpha_i = w_i \nu_i^{-2} = w_i \Big(\sum_{j=1}^{J} |A_{ij}|^2 \Big)^{-1}$$

and $\beta_j = s_j^{-1}$, *for each* i *and* j. *Then the simultaneous algorithm with the iterative step*

$$x_j^k = x_j^{k-1} + \lambda_k \sum_{i=1}^{I} \left(\frac{w_i \overline{A_{ij}} (b_i - (Ax^{k-1})_i)}{s_j \nu_i^2} \right), \tag{27.31}$$

converges to the minimizer of the function

$$\sum_{i=1}^{I} \left| \frac{w_i (b_i - (Ax)_i)}{\nu_i} \right|^2$$

for which the function

$$\sum_{j=1}^{J} s_j |x_j - x_j^0|^2$$

is minimized.

For $w_i = 1$, this is the CARP1 algorithm of [152] (see also [115, 85, 86]). The simultaneous DROP algorithm of [83] requires only that the weights

w_i be positive, but dividing each w_i by their maximum, $\max_i\{w_i\}$, while multiplying each λ_k by the same maximum, gives weights in the interval $(0, 1]$. For convergence of their algorithm, we need to replace the condition $\lambda_k \leq 2 - \epsilon$ with $\lambda_k \leq \frac{2-\epsilon}{\max_i\{w_i\}}$.

The denominator in CAV is

$$\sum_{t=1}^{J} s_t |A_{it}|^2,$$

while that in CARP1 is

$$s_j \sum_{t=1}^{J} |A_{it}|^2.$$

It was reported in [152] that the two methods differed only slightly in the simulated cases studied.

27.9 Block-iterative Algorithms

The methods discussed in the previous section are *simultaneous*, that is, all the equations are employed at each step of the iteration. We turn now to *block-iterative methods*, which employ only some of the equations at each step. When the parameters are appropriately chosen, block-iterative methods can be significantly faster than simultaneous ones.

27.9.1 The Block-Iterative Landweber Algorithm

For a given set of blocks, the block-iterative Landweber algorithm has the following iterative step: with $n = k(\mathrm{mod}\, N)$,

$$x^k = x^{k-1} + \gamma_n A_n^\dagger (b^n - A_n x^{k-1}). \tag{27.32}$$

The sequence $\{x^k\}$ converges to the solution of $Ax = b$ that minimizes $\|x - x^0\|_2$, whenever the system $Ax = b$ has solutions, provided that the parameters γ_n satisfy the inequalities $0 < \gamma_n < 1/L_n$. This follows from Theorem 27.7 by replacing the matrices A_n with $\sqrt{\gamma_n} A_n$ and the vectors b^n with $\sqrt{\gamma_n} b^n$.

If the rows of the matrices A_n are normalized to have length one, then we know that $L_n \leq s_n$. Therefore, we can use parameters γ_n that satisfy

$$0 < \gamma_n \leq \left(s_n \sum_{j=1}^{J} |A_{ij}|^2 \right)^{-1}, \tag{27.33}$$

for each $i \in B_n$.

27.9.2 The BICAV Algorithm

We can extend the block-iterative Landweber algorithm as follows: let $n = k(\mathrm{mod}\, N)$ and

$$x^k = x^{k-1} + \lambda_k \sum_{i \in B_n} \gamma_i (b_i - (Ax^{k-1})_i) A_i^\dagger. \tag{27.34}$$

It follows from Theorem 27.1 that, in the consistent case, the sequence $\{x^k\}$ converges to the solution of $Ax = b$ that minimizes $\|x - x^0\|$, provided that, for each n and each $i \in B_n$, we have

$$\gamma_i \leq \Big(\sum_{j=1}^J s_{nj} |A_{ij}|^2 \Big)^{-1}.$$

The BICAV algorithm [86] uses

$$\gamma_i = \Big(\sum_{j=1}^J s_{nj} |A_{ij}|^2 \Big)^{-1}.$$

The iterative step of BICAV is

$$x^k = x^{k-1} + \lambda_k \sum_{i \in B_n} \left(\frac{b_i - (Ax^{k-1})_i}{\sum_{t=1}^J s_{nt} |A_{it}|^2} \right) A_i^\dagger. \tag{27.35}$$

27.9.3 A Block-Iterative CARP1

The obvious way to obtain a block-iterative version of CARP1 would be to replace the denominator term

$$s_j \sum_{t=1}^J |A_{it}|^2$$

with

$$s_{nj} \sum_{t=1}^J |A_{it}|^2.$$

However, this is problematic, since we cannot redefine the vector of unknowns using $z_j = x_j \sqrt{s_{nj}}$, since this varies with n. In [83], this issue is resolved by taking τ_j to be not less than the maximum of the s_{nj}, and using the denominator

$$\tau_j \sum_{t=1}^J |A_{it}|^2 = \tau_j \nu_i^2.$$

A similar device is used in [175] to obtain a convergent block-iterative version of SART. The iterative step of DROP is

$$x_j^k = x_j^{k-1} + \lambda_k \sum_{i \in B_n} \left(\overline{A_{ij}} \frac{(b_i - (Ax^{k-1})_i)}{\tau_j \nu_i^2} \right). \tag{27.36}$$

Convergence of the DROP (*diagonally-relaxed orthogonal projection*) iteration follows from their Theorem 11. We obtain convergence as a corollary of our previous results.

The change of variables is $z_j = x_j \sqrt{\tau_j}$, for each j. Using our eigenvalue bounds, it is easy to show that the matrices C_n with entries

$$(C_n)_{ij} = \left(\frac{A_{ij}}{\sqrt{\tau_j} \nu_i} \right),$$

for all $i \in B_n$ and all j, have $\rho(C_n^\dagger C_n) \leq 1$. The resulting iterative scheme, which is equivalent to Equation (27.36), then converges, whenever $Ax = b$ is consistent, to the solution minimizing the proximity function

$$\sum_{i=1}^{I} \left| \frac{b_i - (Ax)_i}{\nu_i} \right|^2$$

for which the function

$$\sum_{j=1}^{J} \tau_j |x_j - x_j^0|^2$$

is minimized.

27.9.4 Using Sparseness

Suppose, for the sake of illustration, that each column of A has s nonzero elements, for some $s < I$, and we let $r = s/I$. Suppose also that the number of members of B_n is $I_n = I/N$ for each n, and that N is not too large. Then s_n is approximately equal to $rI_n = s/N$. On the other hand, unless A_n has only zero entries, we know that $s_n \geq 1$. Therefore, it is no help to select N for which $s/N < 1$. For a given degree of sparseness s we need not select N greater than s. The more sparse the matrix A, the fewer blocks we need to gain the maximum advantage from the rescaling, and the more we can benefit from parallelization in the calculations at each step of the algorithm in Equation (27.23).

27.10 Exercises

Ex. 27.3 *Prove Lemma 27.1.*

Ex. 27.4 (Computer Problem) *Compare the speed of convergence of the ART and Cimmino algorithms.*

Ex. 27.5 (Computer Problem) *By generating sparse matrices of various sizes, test the accuracy of the estimates of the largest singular-value given above.*

Chapter 28

Appendix: List-Mode Reconstruction in PET

28.1 Chapter Summary

In modern PET scanners the number of pairs of detectors, and therefore, the number of potential lines of response (LOR), often exceeds the number of detections; the count recorded at any single i is typically one or zero. It makes sense, therefore, to record the data as a list of those LOR corresponding to a detection; this is list-mode data.

28.2 Why List-Mode Processing?

In PET the radionuclide emits individual positrons, which travel, on average, between 4 mm and 2.5 cm (depending on their kinetic energy) before encountering an electron. The resulting annihilation releases two gamma-ray photons that then proceed in essentially opposite directions. Detection in the PET case means the recording of two photons at nearly the same time at two different detectors. The locations of these two detectors then provide the end points of the line segment passing, more or less, through the site of the original positron emission. Therefore, each possible

pair of detectors determines a line of response. When a LOR is recorded, it is assumed that a positron was emitted somewhere along that line.

28.3 Correcting for Attenuation in PET

In SPECT attenuation correction is performed by modifying the probabilities P_{ij}. In PET the situation is at once simpler and more involved.

Let a given LOR be parameterized by the variable s, with $s = 0$ and $s = c$ denoting the two ends, and c the distance from one end to the other. For a fixed value $s = s_0$, let $P(s)$ be the probability of reaching s for a photon resulting from an emission at s_0. For small $\Delta s > 0$ the probability that a photon that reached s is absorbed in the interval $[s, s + \Delta s]$ is approximately $\mu(s)\Delta s$, where $\mu(s) \geq 0$ is the photon attenuation density at s. Then $P(s + \Delta s) \approx P(s)[1 - \mu(s)\Delta s]$, so that

$$P(s + \Delta s) - P(s) \approx -P(s)\mu(s)\Delta s.$$

Dividing by Δs and letting Δs go to zero, we get

$$P'(s) = -P(s)\mu(s).$$

It follows that

$$P(s) = e^{-\int_{s_0}^{s} \mu(t)dt}.$$

The probability that the photon will reach $s = c$ and be detected is then

$$P(c) = e^{-\int_{s_0}^{c} \mu(t)dt}.$$

Similarly, we find that the probability that a photon will succeed in reaching $s = 0$ from s_0 is

$$P(0) = e^{-\int_{0}^{s_0} \mu(t)dt}.$$

Since having one photon reach $s = 0$ and the other reach $s = c$ are independent events, their probabilities multiply, so that the probability that both photons reach their destinations and a coincident detection is recorded for this LOR is

$$e^{-\int_{0}^{c} \mu(t)dt}.$$

The expected number of coincident detections along the LOR is then proportional to

$$\int_0^c f(s)e^{-\int_0^c \mu(t)dt}ds = e^{-\int_0^c \mu(t)dt} \int_0^c f(s)ds, \tag{28.1}$$

where $f(s)$ is the intensity of radionuclide at s.

For each LOR i and each pixel or voxel j, let A_{ij} be the *geometric probability* that an emission at j will result in two photons traveling along the LOR i. The probability A_{ij} is unrelated to the attenuation presented by the body of the patient. Then the probability that an emission at j will result in the LOR i being added to the list is

$$P_{ij} = a_i A_{ij},$$

where

$$a_i = e^{-\int_i \mu(s)ds},$$

and the integral is the line integral along the line segment associated with the LOR i. We then perform attenuation correction by using the probabilities P_{ij} in the reconstruction.

Note that, if the number I of potential LOR is not too large and the entries of the data vector y are not simply zero or one, we might correct for attenuation by replacing each y_i with y_i/a_i, which is approximately the count we would have seen for the LOR i if there had been no attenuation. However, in the more typical case of large I and zero or one values for the y_i, this approach does not make much sense. The effect of attenuation now is to prevent certain i from being recorded, not to diminish the values of the positive y_i of the LOR that were recorded. Therefore, at least in theory, it makes more sense to correct for attenuation by using the P_{ij}. There is an additional complication, though.

In list-mode processing, I, the number of potential LOR, is much larger than the size of the list. To employ the EMML algorithm or one of its block-iterative variants, we need to calculate the probabilities associated with those LOR on the list, but it is costly to do this for all the potential LOR; we do need to compute the sensitivities, or probabilities of detection, for each pixel, however. If we consider only the geometry of the scanner, calculating the sensitivities for each pixel is not difficult and can be done once and used repeatedly; it is much more problematic if we must include the patient-specific attenuation. For this reason, it makes sense, practically speaking, to correct for attenuation in list-mode PET by replacing y_i with y_i/a_i for those y_i equal to one. The reconstruction is probably much the same, either way.

28.4 Modeling the Possible LOR

We can model the potential LOR simply as pairs of detectors, so that I, the number of potential LOR, is very large, but finite, and finite probability

vectors, rather than probability density functions, suffice in forming the likelihood function. The EMML algorithm applies directly to this list-mode model. This is the approach adopted by Huesman *et al.* [173].

Alternatively, one can assume that the end-point coordinates form a continuum, so that the set of potential LOR is uncountably infinite. Now we need probability density functions to form the likelihood function. This method, adopted by Parra and Barrett [216], makes the application of the EMML algorithm more complicated, as discussed in [61].

28.5 EMML: The Finite LOR Model

In this section we discuss the EMML iterative algorithm for list-mode reconstruction based on the finite model.

Let the list of recorded LOR be $\{i_1, ..., i_M\}$ and let

$$Q_{mj} = P_{i_m,j},$$

for $m = 1, ..., M$. Since the values of the y_i are typically zero or one, the i_m are typically distinct, but this is not essential here. The EMML iteration becomes

$$x_j^{k+1} = x_j^k s_j^{-1} \sum_{m=1}^{M} Q_{mj} \Big(\frac{1}{(Qx^k)_m} \Big). \tag{28.2}$$

Note that we still need to use the sensitivity values

$$s_j = \sum_{i=1}^{I} P_{ij},$$

which are the probabilities of detection. However, for imaging the radionuclide we do not need to calculate the s_j by first determining each of the P_{ij}; we need only that the $s_j > \sum_{m=1}^{M} Q_{mj}$ for each j and that the relative values of the various s_j be reasonably accurate. For quantitation, though, accurate absolute values of the s_j are needed.

28.6 List-mode RBI-EMML

We turn now to the block-iterative versions of EMML. For $n = 1, ..., N$ let C_n consist of all indices m such that the LOR i_m on the list is also in

B_n. The list-mode BI-EMML (LMBI-EMML) has the iterative step

$$x_j^k = (1 - \gamma_n \delta_j s_{nj}) x_j^{k-1} + x_j^k \gamma_n \delta_j \sum_{m \in C_n} P_{ij} \Big(\frac{1}{(Qx^k)_m} \Big), \tag{28.3}$$

with $\gamma > 0$ chosen so that

$$s_{nj} \delta_j \gamma_n \leq 1.$$

When we select $\delta_j = s_j^{-1}$, we must then have $\gamma_n \leq \mu_n^{-1}$. When we have $\delta_j = 1$, we need $\gamma_n \leq m_n^{-1}$. Generally speaking, the larger the γ_n the faster the convergence. The *rescaled* LMBI-EMML (LMRBI-EMML) uses the largest values of γ_n consistent with these constraints.

Note that, as previously, we need s_j and now we also need s_{nj}. As before, though, we do not need to specify each of the P_{ij} to obtain reasonable choices for these values.

28.7 The Row-action LMRBI-EMML: LMEMART

The row-action or *event-by-event* version of the RBI-EMML algorithm, the LMEMART, is a special case of the LMRBI-EMML in which, for $m = 1, ..., M$, each LOR i_m on the list forms its own block or subset, denoted C_m. Another way to say this is that we choose the original blocks B_n so that no B_n contains more than one i_m. For clarity, we shall assume that the blocks B_n are chosen so that $B_m = \{i_m\}$ and $C_m = \{m\}$, for $m = 1, ..., M$. We then let B_{M+1} consist of all the i not equal to some I_m on the list, and $N = M + 1$. Therefore, for $n = 1, ..., M$, we have

$$s_{nj} = Q_{nj}.$$

In the LMEMART each iteration employs a single member of the list and we cycle through the list repeatedly. The iteration index is now $m = 1, ..., M$, with $m = m(k) = k(\text{mod}\, M) + 1$.

The LMEMART has the iterative step

$$x_j^{k+1} = (1 - \gamma_m \delta_j Q_{mj}) x_j^k + x_j^k \gamma_m \delta_j Q_{mj} \Big(\frac{1}{(Qx^k)_m} \Big), \tag{28.4}$$

with $Q_{mj} \delta_j \gamma_m \leq 1$.

28.8 EMML: The Continuous LOR Model

When the end points of the potential LOR are allowed to take on values in a continuum, the likelihood function involves probability density functions, rather than finite probabilities. This poses a difficulty, in that the values of probability density functions can be any non-negative real number; only their integrals are required to be one. As a result, the convergence theory for the EMML algorithm and its various block-iterative versions does not apply unchanged.

For each pixel index j, let $f_j(\cdot)$ be the probability density function (pdf) whose domain is the (uncountably infinite) set of potential LOR with the property that the probability that an emission at j results in an LOR from the set S being recorded is the integral of f_j over S. With x_j the expected number of emissions from j during the scanning time, and

$$x_+ = \sum_{j=1}^{J} x_j,$$

the probability that an emission came from j, given that an emission has happened, is x_j/x_+. Therefore, the probability that an LOR in the set S will be recorded, given that an emission has happened, is the integral over S of the pdf

$$f(\cdot) = \frac{1}{x_+} \sum_{j=1}^{J} x_j f_j(\cdot).$$

For each j let d_j be the probability that an emission from j will be detected, and let

$$d = \frac{1}{x_+} \sum_{j=1}^{J} x_j d_j$$

be the probability that an emission will be detected.

The number of items on the list, M, is also a random variable, which we model as having a Poisson distribution with mean value dx_+. Therefore, the probability of M is

$$p(M) = \exp(-x_+ d)(x_+ d)^M / M!.$$

Given the list of recorded LOR, the likelihood function is then

$$L(x) = p(M) \prod_{m=1}^{M} f(i_m),$$

and the log likelihood function to be maximized is

$$LL(x) = -x_+ d + \sum_{m=1}^{M} \log(Px)_m,$$

where the matrix P has entries

$$P_{mj} = f_j(i_m).$$

Note that

$$(Px)_m = \sum_{j=1}^{J} P_{mj} x_j,$$

so that

$$\sum_{m=1}^{M} (Px)_m = \sum_{j=1}^{J} \Big(\sum_{m=1}^{M} P_{mj}) x_j = \sum_{j=1}^{J} c_j x_j,$$

for

$$c_j = \sum_{m=1}^{M} P_{mj}.$$

Maximizing the log likelihood function is equivalent to minimizing

$$KL(u, Px) - \sum_{m=1}^{M} (Px)_m + x_+ d + \text{constants},$$

where u is the vector whose entries are all one, and therefore equivalent to minimizing

$$F(x) = KL(u, Px) + \sum_{j=1}^{J} (d_j - c_j) x_j.$$

The EMML algorithm itself will minimize only $KL(u, Px)$. The basic problem now is that we have values of probability density functions and the quantities c_j, which can be any positive real numbers, are unrelated to the detectability or sensitivity d_j.

It was shown in [61] that the EMML algorithm can be modified to provide a convergent iterative method for minimizing $F(x)$. This modified EMML algorithm has the iterative step

$$x_j^{k+1} = x_j^k d_j^{-1} \sum_{m=1}^{M} \Big(\frac{1}{(Px^k)_m} \Big).$$

For the finite model, as in [173], this is just the usual EMML and convergence follows from known results, but for the continuous model, as in [216],

this iterative scheme falls outside the EMML framework and convergence needed to be established, as in [61].

Just as the EMML algorithm must be modified before it can be applied to the continuous model, we must adapt the block-iterative versions as well; see [61] for details.

Chapter 29

Appendix: A Little Optimization

29.1 Chapter Summary

In our discussion of both transmission and emission tomography we saw that discretization leads to systems of linear equations to be solved for the vectorized image x. Typically, these systems are quite large, the measured data is noisy, and there will be no non-negative x satisfying the system exactly. In such cases, one can turn to optimization, and calculate a non-negatively constrained least-squares solution, with or without a penalty term.

29.2 Image Reconstruction Through Optimization

In the stochastic approach to emission tomography, we maximize the likelihood function with respect to the unknown image vector x. Here again,

optimization plays a role. It is reasonable, therefore, to take a brief look at the theory of optimization, particularly constrained optimization. In this chapter we discuss optimization with equality constraints and the area known as *convex programming* (CP).

29.3 Eigenvalues and Eigenvectors Through Optimization

Let B be any real I by J matrix. We want to find the maximum value of the ratio $||Bx||/||x||$, over all non-zero vectors x. If $\hat{x}$ solves this problem, so does $c\hat{x}$ for every non-zero real number c; therefore, we may and do constrain the vectors x to have $||x|| = 1$.

We reformulate the problem as follows: maximize $f(x) = ||Bx||^2$, subject to $g(x) = ||x||^2 = 1$. Our approach will be to use the method of *Lagrange multipliers*. Suppose that $\hat{x}$ is a solution and S is the level surface of the function $f(x)$ containing the vector $\hat{x}$, that is,

$$S = \{x | f(x) = f(\hat{x})\}.$$

The gradient of $f(x)$ at $\hat{x}$ is a vector normal to S at $\hat{x}$. Now let U be the unit surface of all x with $||x|| = 1$. We claim that S and U must be tangent at $x = \hat{x}$. If that is not the case, then U cuts through S, making it possible to move from one side of S to the other side of S, while remaining on the surface U. Therefore, we would be able to move along U to another vector x with $f(x) > f(\hat{x})$, which cannot happen.

Since the two surfaces are tangent at $x = \hat{x}$, their gradients are parallel, so that

$$\nabla f(\hat{x}) = \alpha \nabla g(\hat{x}),$$

for some constant α. Equivalently,

$$\nabla f(\hat{x}) + (-\alpha)\nabla g(\hat{x}) = 0.$$

The main idea of the Lagrange-multiplier method is to define the Lagrangian as

$$L(x; \lambda) = f(x) + \lambda g(x),$$

so that, for some value of the parameter λ the gradient of $L(x; \lambda)$ is zero; here $\lambda = -\alpha$ works.

The *Lagrangian* for this problem is

$$L(x, \lambda) = f(x) + \lambda g(x) = ||Bx||^2 + \lambda ||x||^2.$$

Therefore, we have

$$2B^TB\hat{x} + 2\lambda\hat{x} = 0,$$

or

$$B^TB\hat{x} = \alpha\hat{x},$$

which tells us that $\hat{x}$ is an *eigenvector* of the matrix B^TB corresponding to the *eigenvalue* α. Since the matrix B^TB is symmetric, all its eigenvalues are real numbers; in fact, B^TB is non-negative definite, so all its eigenvalues are non-negative.

Since

$$||B\hat{x}||^2 = \hat{x}^TB^TB\hat{x} = \alpha\hat{x}^T\hat{x} = \alpha||\hat{x}||^2 = \alpha,$$

we see that the largest value of $||Bx||^2$, subject to $||x|| = 1$, must be α. So α is the largest eigenvalue of the matrix B^TB and $\hat{x}$ is an associated eigenvector.

The largest eigenvalue of B^TB is also the largest eigenvalue of the matrix BB^T and is denoted $\rho(B^TB) = \rho(BB^T)$, and called the *spectral radius* of B^TB. We can therefore write

$$||Bz||^2 \leq \rho(B^TB)||z||^2, \tag{29.1}$$

for all vectors z.

29.4 Convex Sets and Convex Functions

A subset C of $\mathbb{R}^J$ is said to be *convex* if, for every collection $c_1, c_2, ..., c_N$ of points in C and all positive constants $a_1, a_2, ..., a_N$ summing to one, the point $a_1c_1 + ... + a_Nc_N$ is again in C. A function $f : \mathbb{R}^J \rightarrow R$ is said to be a *convex function* on the convex set C if, for all such combinations as above, we have

$$f(a_1c_1 + ... + a_Nc_N) \leq a_1f(c_1) + ... + a_Nf(c_N).$$

The function $f(x) = ||Ax - b||^2$ is convex on $C = \mathbb{R}^J$ and the function $f(x) = KL(b, Ax)$ is convex on the set C of non-negative x in $\mathbb{R}^J$.

29.5 The Convex Programming Problem

Let f and g_i, $i = 1, ..., I$, be convex functions defined on a non-empty closed convex subset C of $\mathbb{R}^J$. The *primal problem* in *convex programming*

(CP) is the following:

$$\text{minimize } f(x), \text{ subject to } g_i(x) \leq 0, \text{ for } i = 1, ..., I. \quad \text{(P)} \tag{29.2}$$

For notational convenience, we define $g(x) = (g_1(x), ..., g_I(x))$. Then (P) becomes

$$\text{minimize } f(x), \text{ subject to } g(x) \leq 0. \quad \text{(P)} \tag{29.3}$$

The *feasible set* for (P) is

$$F = \{x | g(x) \leq 0\}. \tag{29.4}$$

Definition 29.1 *The problem (P) is said to be* consistent *if F is not empty, and* super-consistent *if there is x in F with $g_i(x) < 0$ for all $i = 1, ..., I$. Such a point x is then called a* Slater point.

Definition 29.2 *The* Lagrangian *for the problem (P) is the function*

$$L(x, \lambda) = f(x) + \sum_{i=1}^{I} \lambda_i g_i(x), \tag{29.5}$$

defined for all x in C and $\lambda \geq 0$.

29.6 A Simple Example

Let us minimize the function $f : \mathbb{R}^2 \to \mathbb{R}$ given by

$$f(x, y) = (x + 1)^2 + y^2,$$

subject to $x \geq 0$ and $y \geq 0$. To get this problem into the form of the CP problem we introduce the functions

$$g_1(x, y) = -x,$$

and

$$g_2(x, y) = -y.$$

The partial derivative of f, with respect to x, is

$$\frac{\partial f}{\partial x}(x, y) = 2(x + 1),$$

and the partial derivative of f, with respect to y, is

$$\frac{\partial f}{\partial y}(x, y) = 2y.$$

If we simply set both partial derivatives to zero, we get $x = -1$ and $y = 0$, which is, of course, the unconstrained minimizing point for f. But this point does not satisfy our constraints.

If we graph the function, we see immediately that the constrained solution is the origin, $x = 0$ and $y = 0$. At this point, we can move up or down without decreasing f, and this is reflected in the fact that the y-partial derivative at $(0, 0)$ is zero. The x-partial derivative at $(0, 0)$ is not zero, however, since, if we move horizontally to the left, the function f decreases. However, we are prevented from moving left by the constraint that $x \geq 0$, so it is not necessary that the x-partial derivative be zero at the solution. We only need to know that if we move to the right, which is permitted by the constraints, the function f increases; the fact that the x-partial derivative is positive at $(0, 0)$ guarantees this.

29.7 The Karush-Kuhn-Tucker Theorem

As we have just seen, at the solution of a CP problem it is not necessarily the case that the partial derivatives all be zero. But what does have to be the case?

The Karush-Kuhn Tucker Theorem gives necessary and sufficient conditions for a vector x^* to be a solution of a super-consistent problem (P).

Theorem 29.1 *Let (P) be super-consistent. Then x^* solves (P) if and only if there is a vector λ^* such that*

- **1)** $\lambda^* \geq 0$*;*
- **2)** $\lambda_i^* g_i(x^*) = 0$*, for all* $i = 1, ..., I$*;*
- **3)** $\nabla f(x^*) + \sum_{i=1}^{I} \lambda_i^* \nabla g_i(x^*) = 0$.

When we optimize subject to an equality constraint the first condition of the KKT Theorem need not hold, that is, the Lagrange multipliers need not be non-negative, and the second condition is automatically true, since the constraints are now $g_i(x) = 0$ for all i.

29.8 Back to our Example

Once again, the problem is to minimize $f(x, y) = (x+1)^2 + y^2$, subject to $g_1(x, y) = -x \leq 0$ and $g_2(x, y) = -y \leq 0$. Applying Condition 3 of the KKT Theorem, we get

$$0 = 2(x + 1) - \lambda_1^*,$$

and

$$0 = 2y - \lambda_2^*.$$

From Condition 2 we know that either $\lambda_1^* = 0$, which can't happen, since then $x = -1$, or $x = 0$; therefore $x = 0$. Also from Condition 2 we know that either $\lambda_2^* = 0$ or $y = 0$; therefore, $y = 0$. We have found the solution to our constrained minimization problem.

29.9 Two More Examples

We illustrate the use of the gradient form of the KKT Theorem with two more examples that appeared in the paper of Driscoll and Fox [117].

29.9.1 A Linear Programming Problem

Minimize $f(x_1, x_2) = 3x_1 + 2x_2$, subject to the constraints $2x_1 + x_2 \geq 100$, $x_1 + x_2 \geq 80$, $x_1 \geq 0$ and $x_2 \geq 0$. We define

$$g_1(x_1, x_2) = 100 - 2x_1 - x_2 \leq 0, \tag{29.6}$$

$$g_2(x_1, x_2) = 80 - x_1 - x_2, \tag{29.7}$$

$$g_3(x_1, x_2) = -x_1, \tag{29.8}$$

and

$$g_4(x_1, x_2) = -x_2. \tag{29.9}$$

The Lagrangian is then

$$L(x, \lambda) = 3x_1 + 2x_2 + \lambda_1(100 - 2x_1 - x_2)$$

$$+\lambda_2(80 - x_1 - x_2) - \lambda_3 x_1 - \lambda_4 x_2.$$

(29.10)

From the KKT Theorem, we know that if there is a solution x^*, then there is $\lambda^* \geq 0$ with

$$f(x^*) = L(x^*, \lambda^*) \leq L(x, \lambda^*),$$

for all x. For notational simplicity, we write λ in place of λ^*.

Taking the partial derivatives of $L(x, \lambda)$ with respect to the variables x_1 and x_2, we get

$$3 - 2\lambda_1 - \lambda_2 - \lambda_3 = 0, \tag{29.11}$$

and

$$2 - \lambda_1 - \lambda_2 - \lambda_4 = 0. \tag{29.12}$$

The complementary slackness conditions are

$$\lambda_1 = 0 \ , \text{ if } 2x_1 + x_2 \neq 100, \tag{29.13}$$

$$\lambda_2 = 0 \ , \text{ if } x_1 + x_2 \neq 80, \tag{29.14}$$

$$\lambda_3 = 0 \ , \text{ if } x_1 \neq 0, \tag{29.15}$$

and

$$\lambda_4 = 0 \ , \text{ if } x_2 \neq 0. \tag{29.16}$$

A little thought reveals that precisely two of the four constraints must be binding. Examining the six cases, we find that the only case satisfying all the conditions of the KKT Theorem is $\lambda_3 = \lambda_4 = 0$. The minimum occurs at $x_1 = 20$ and $x_2 = 60$ and the minimum value is $f(20, 60) = 180$.

29.9.2 A Nonlinear Convex Programming Problem

Minimize the function

$$f(x_1, x_2) = (x_1 - 14)^2 + (x_2 - 11)^2,$$

subject to

$$g_1(x_1, x_2) = (x_1 - 11)^2 + (x_2 - 13)^2 - 49 \leq 0,$$

and

$$g_2(x_1, x_2) = x_1 + x_2 - 19 \leq 0.$$

The Lagrangian is then

$$L(x,\lambda) = (x_1 - 14)^2 + (x_2 - 11)^2 +$$

$$\lambda_1\Big((x_1 - 11)^2 + (x_2 - 13)^2 - 49\Big) + \lambda_2\Big(x_1 + x_2 - 19\Big). \tag{29.17}$$

Again, we write λ in place of λ^*. Setting the partial derivatives, with respect to x_1 and x_2, to zero, we get the KKT equations

$$2x_1 - 28 + 2\lambda_1 x_1 - 22\lambda_1 + \lambda_2 = 0, \tag{29.18}$$

and

$$2x_2 - 22 + 2\lambda_1 x_2 - 26\lambda_1 + \lambda_2 = 0. \tag{29.19}$$

The complementary slackness conditions are

$$\lambda_1 = 0\,, \quad \text{if } (x_1 - 11)^2 + (x_2 - 13)^2 \neq 49, \tag{29.20}$$

and

$$\lambda_2 = 0\,, \quad \text{if } x_1 + x_2 \neq 19. \tag{29.21}$$

There are four cases to consider. First, if neither constraint is binding, the KKT equations have solution $x_1 = 14$ and $x_2 = 11$, which is not feasible. If only the first constraint is binding, we obtain two solutions, neither feasible. If only the second constraint is binding, we obtain $x_1^* = 11$, $x_2^* = 8$, and $\lambda_2 = 6$. This is the optimal solution. If both constraints are binding, we obtain, with a bit of calculation, two solutions, neither feasible. The minimum value is $f(11, 8) = 18$, and the sensitivity vector is $\lambda^* = (0, 6)$.

29.10 Non-Negatively Constrained Least-Squares

If there is no solution to a system of linear equations $Ax = b$, then we may seek a *least-squares* "solution" , which is a minimizer of the function

$$f(x) = \sum_{i=1}^{I} \Big((\sum_{m=1}^{J} A_{im} x_m) - b_i \Big)^2 = ||Ax - b||^2.$$

The partial derivative of $f(x)$ with respect to the variable x_j is

$$\frac{\partial f}{\partial x_j}(x) = 2 \sum_{i=1}^{I} A_{ij} \Big((\sum_{m=1}^{J} A_{im} x_m) - b_i \Big).$$

Setting the gradient equal to zero, we find that to get a least-squares solution we must solve the system of equations

$$A^T(Ax - b) = 0.$$

Now we consider what happens when the additional constraints $x_j \geq 0$ are imposed.

This problem fits into the CP framework, when we define

$$g_j(x) = -x_j,$$

for each j. Let $\hat{x}$ be a least-squares solution. According to the KKT Theorem, for those values of j for which $\hat{x}_j$ is not zero we have $\lambda_j^* = 0$ and $\frac{\partial f}{\partial x_j}(\hat{x}) = 0$. Therefore, if $\hat{x}_j \neq 0$,

$$0 = \sum_{i=1}^{I} A_{ij}\Big((\sum_{m=1}^{J} A_{im}\hat{x}_m) - b_i\Big).$$

Let Q be the matrix obtained from A by deleting rows j for which $\hat{x}_j = 0$. Then we can write

$$Q^T(A\hat{x} - b) = 0.$$

If Q has at least I columns and has full rank, then Q^T is a one-to-one linear transformation, which implies that $A\hat{x} = b$. Therefore, when there is no non-negative solution of $Ax = b$, Q must have fewer than I columns, which means that $\hat{x}$ has fewer than I non-zero entries. This is the proof of Theorem 24.1.

This result has some practical implications in medical image reconstruction. In the hope of improving the resolution of the reconstructed image, we may be tempted to take J, the number of pixels, larger than I, the number of equations arising from photon counts or line integrals. Since the vector b consists of measured data, it is noisy and there may well not be a non-negative solution of $Ax = b$. As a result, the image obtained by non-negatively constrained least-squares will have at most $I - 1$ non-zero entries; many of the pixels will be zero and they will be scattered throughout the image, making it unusable for diagnosis. The reconstructed images resemble stars in a night sky, and, as a result, the theorem is sometimes described as the "night sky" theorem.

This "night sky" phenomenon is not restricted to least squares. The same thing happens with methods based on the Kullback-Leibler distance, such as MART, EMML and SMART.

29.11 The EMML Algorithm

Maximizing the likelihood function in SPECT is equivalent to minimizing the KL distance $KL(b, Ax)$ over non-negative vectors x, where b is the vector of photon counts at the detectors and A the matrix of detection probabilities. With $f(x) = KL(b, Ax)$ and $g_j(x) = -x_j$, the problem becomes a CP problem. We have

$$\frac{\partial f}{\partial x_j}(x) = \sum_{i=1}^{I} A_{ij}\Big(1 - b_i/(Ax)_i\Big),$$

where

$$(Ax)_i = \sum_{m=1}^{J} A_{im}x_m.$$

Let $\hat{x}$ be the solution. According to the Karush-Kuhn-Tucker Theorem 29.1, one of two things are possible: for each j either 1): $\hat{x}_j = 0$ or 2): both $\lambda_j^* = 0$ and, consequently,

$$\frac{\partial f}{\partial x_j}(\hat{x}) = 0.$$

Therefore, for all values of the index j we have

$$0 = \hat{x}_j \sum_{i=1}^{I} A_{ij}\Big(1 - b_i/(A\hat{x})_i\Big),$$

or, equivalently,

$$\hat{x}_j = \hat{x}_j s_j^{-1} \sum_{i=1}^{I} A_{ij}\Big(b_i/(A\hat{x})_i\Big),$$

where $s_j = \sum_{i=1}^{I} A_{ij}$.

This suggests an iterative optimization algorithm whereby we insert the current value of the vector, call it x^k, into the right side of the last equation, and call the resulting vector the next iterate, x^{k+1}. For simplicity, we assume $s_j = 1$. Then the iteration becomes

$$x_j^{k+1} = x_j^k\Big(\sum_{i=1}^{I} A_{ij}(b_i/(Ax^k)_i)\Big). \tag{29.22}$$

This is the EMML iterative algorithm.

29.12 The Simultaneous MART Algorithm

The MART algorithm has the following iterative step:

$$x_j^{k+1} = x_j^k \Big(b_i / (Ax^k)_i \Big)^{A_{ij}},$$

where $i = k(\mathrm{mod}\, I) + 1$. The MART uses only one equation at each step. The simultaneous MART (SMART) uses all the equations at each step. Assuming once again that $s_j = 1$ for all j, the iterative step of the SMART is

$$x_j^{k+1} = x_j^k \exp \Big(\sum_{i=1}^{I} A_{ij} \log(b_i / (Ax^k)_i) \Big). \tag{29.23}$$

The SMART is clearly closely related to the EMML algorithm, with subtle differences, namely the exponentiation and the logarithm. As we shall show in the next chapter, the SMART algorithm minimizes the function $KL(Ax, b)$, while the EMML minimizes $KL(b, Ax)$.

Chapter 30

Appendix: Geometric Programming and the MART

30.1 Chapter Summary

In this chapter we present yet another application of the MART. Geometric Programming (GP) involves the minimization of functions of a special type, known as posynomials. The first systematic treatment of geometric programming appeared in the book [121], by Duffin, Peterson and Zener, the founders of geometric programming. As we shall see, the Generalized Arithmetic-Geometric Mean Inequality plays an important role in the theoretical treatment of geometric programming, particularly in the development of the dual GP (DGP) problem. The MART is then used to solve the DGP.

30.2 An Example of a GP Problem

The following optimization problem was presented originally by Duffin, *et al.* [121] and discussed by Peressini *et al.* in [218]. It illustrates well the type of problem considered in geometric programming. Suppose that 400 cubic yards of gravel must be ferried across a river in an open box of length t_1, width t_2 and height t_3. Each round-trip cost ten cents. The sides and the bottom of the box cost 10 dollars per square yard to build, while the ends of the box cost twenty dollars per square yard. The box will have no salvage value after it has been used. Determine the dimensions of the box that minimize the total cost.

With $t = (t_1, t_2, t_3)$, the cost function is

$$g(t) = \frac{40}{t_1 t_2 t_3} + 20t_1t_3 + 10t_1t_2 + 40t_2t_3, \tag{30.1}$$

which is to be minimized over $t_j > 0$, for $j = 1, 2, 3$. The function $g(t)$ is an example of a posynomial.

30.3 The Generalized AGM Inequality

The generalized arithmetic-geometric mean inequality will play a prominent role in solving the GP problem.

Suppose that $x_1, ..., x_N$ are positive numbers. Let $a_1, ..., a_N$ be positive numbers that sum to one. Then the *Generalized AGM Inequality* (GAGM Inequality) is

$$x_1^{a_1} x_2^{a_2} \cdots x_N^{a_N} \leq a_1x_1 + a_2x_2 + ... + a_Nx_N, \tag{30.2}$$

with equality if and only if $x_1 = x_2 = ... = x_N$. We can prove this using the convexity of the function $-\log x$.

A function $f(x)$ is said to be *convex* over an interval (a, b) if

$$f(a_1t_1 + a_2t_2 + ... + a_Nt_N) \leq a_1f(t_1) + a_2f(t_2) + ... + a_Nf(t_N),$$

for all positive integers N, all a_n as above, and all real numbers t_n in (a, b). If the function $f(x)$ is twice differentiable on (a, b), then $f(x)$ is convex over (a, b) if and only if the second derivative of $f(x)$ is non-negative on (a, b). For example, the function $f(x) = -\log x$ is convex on the positive x-axis. The GAGM Inequality follows immediately.

30.4 Posynomials and the GP Problem

Functions $g(t)$ of the form

$$g(t) = \sum_{i=1}^{n} c_i \Big(\prod_{j=1}^{m} t_j^{a_{ij}} \Big), \tag{30.3}$$

with $t = (t_1, ..., t_m)$, the $t_j > 0$, $c_i > 0$ and a_{ij} real, are called *posynomials*. The *geometric programming problem*, denoted (GP), is to minimize a given posynomial over positive t. In order for the minimum to be greater than zero, we need some of the a_{ij} to be negative.

We denote by $u_i(t)$ the function

$$u_i(t) = c_i \prod_{j=1}^{m} t_j^{a_{ij}}, \tag{30.4}$$

so that

$$g(t) = \sum_{i=1}^{n} u_i(t). \tag{30.5}$$

For any choice of $\delta_i > 0$, $i = 1, ..., n$, with

$$\sum_{i=1}^{n} \delta_i = 1,$$

we have

$$g(t) = \sum_{i=1}^{n} \delta_i \Big(\frac{u_i(t)}{\delta_i} \Big). \tag{30.6}$$

Applying the Generalized Arithmetic-Geometric Mean (GAGM) Inequality, we have

$$g(t) \geq \prod_{i=1}^{n} \Big(\frac{u_i(t)}{\delta_i} \Big)^{\delta_i}. \tag{30.7}$$

Therefore,

$$g(t) \geq \prod_{i=1}^{n} \Big(\frac{c_i}{\delta_i} \Big)^{\delta_i} \left(\prod_{i=1}^{n} \prod_{j=1}^{m} t_j^{a_{ij}\delta_i} \right), \tag{30.8}$$

or

$$g(t) \geq \prod_{i=1}^{n} \Big(\frac{c_i}{\delta_i}\Big)^{\delta_i} \Big(\prod_{j=1}^{m} t_j^{\sum_{i=1}^{n} a_{ij}\delta_i}\Big), \tag{30.9}$$

Suppose that we can find $\delta_i > 0$ with

$$\sum_{i=1}^{n} a_{ij}\delta_i = 0, \tag{30.10}$$

for each j. Then the inequality in (30.9) becomes

$$g(t) \geq v(\delta), \tag{30.11}$$

for

$$v(\delta) = \prod_{i=1}^{n} \Big(\frac{c_i}{\delta_i}\Big)^{\delta_i}. \tag{30.12}$$

30.5 The Dual GP Problem

The *dual geometric programming problem*, denoted (DGP), is to maximize the function $v(\delta)$, over all *feasible* $\delta = (\delta_1, ..., \delta_n)$, that is, all positive δ for which

$$\sum_{i=1}^{n} \delta_i = 1, \tag{30.13}$$

and

$$\sum_{i=1}^{n} a_{ij}\delta_i = 0, \tag{30.14}$$

for each $j = 1, ..., m$. Clearly, we have

$$g(t) \geq v(\delta), \tag{30.15}$$

for any positive t and feasible δ. Of course, there may be no feasible δ, in which case (DGP) is said to be *inconsistent*.

As we have seen, the inequality in (30.15) is based on the GAGM Inequality. We have equality in the GAGM Inequality if and only if the terms

in the arithmetic mean are all equal. In this case, this says that there is a constant λ such that

$$\frac{u_i(t)}{\delta_i} = \lambda, \tag{30.16}$$

for each $i = 1, ..., n$. Using the fact that the δ_i sum to one, it follows that

$$\lambda = \sum_{i=1}^{n} u_i(t) = g(t), \tag{30.17}$$

and

$$\delta_i = \frac{u_i(t)}{g(t)}, \tag{30.18}$$

for each $i = 1, ..., n$. As the theorem below asserts, if t^* is positive and minimizes $g(t)$, then δ^*, the associated δ from Equation (30.18), is feasible and solves (DGP). Since we have equality in the GAGM Inequality now, we have

$$g(t^*) = v(\delta^*).$$

The main theorem in geometric programming is the following.

Theorem 30.1 *If $t^* > 0$ minimizes $g(t)$, then (DGP) is consistent. In addition, the choice*

$$\delta_i^* = \frac{u_i(t^*)}{g(t^*)} \tag{30.19}$$

is feasible and solves (DGP). Finally,

$$g(t^*) = v(\delta^*); \tag{30.20}$$

that is, there is no duality gap*.*

Proof: We have

$$\frac{\partial u_i}{\partial t_j}(t^*) = \frac{a_{ij} u_i(t^*)}{t_j^*}, \tag{30.21}$$

so that

$$t_j^* \frac{\partial u_i}{\partial t_j}(t^*) = a_{ij} u_i(t^*), \tag{30.22}$$

for each $j = 1, ..., m$. Since t^* minimizes $g(t)$, we have

$$0 = \frac{\partial g}{\partial t_j}(t^*) = \sum_{i=1}^{n} \frac{\partial u_i}{\partial t_j}(t^*), \tag{30.23}$$

so that, from Equation (30.22), we have

$$0 = \sum_{i=1}^{n} a_{ij} u_i(t^*), \tag{30.24}$$

for each $j = 1, ..., m$. It follows that δ^* is feasible. Since we have equality in the GAGM Inequality, we know

$$g(t^*) = v(\delta^*). \tag{30.25}$$

Therefore, δ^* solves (DGP). This completes the proof. ∎

30.6 Solving the GP Problem

The theorem suggests how we might go about solving (GP). First, we try to find a feasible δ^* that maximizes $v(\delta)$. This means we have to find a positive solution to the system of $m + 1$ linear equations in n unknowns, given by

$$\sum_{i=1}^{n} \delta_i = 1, \tag{30.26}$$

and

$$\sum_{i=1}^{n} a_{ij} \delta_i = 0, \tag{30.27}$$

for $j = 1, ..., m$, such that $v(\delta)$ is maximized. As we shall see, the *multiplicative algebraic reconstruction technique* (MART) is an iterative procedure that we can use to find such δ. If there is no such vector, then (GP) has no minimizer. Once the desired δ^* has been found, we set

$$\delta_i^* = \frac{u_i(t^*)}{v(\delta^*)}, \tag{30.28}$$

for each $i = 1, ..., n$, and then solve for the entries of t^*. This last step can be simplified by taking logs; then we have a system of linear equations to solve for the values $\log t_j^*$.

30.7 Solving the DGP Problem

The iterative multiplicative algebraic reconstruction technique MART can be used to minimize the function $v(\delta)$, subject to linear equality constraints, provided that the matrix involved has nonnegative entries. We cannot apply the MART yet, because the matrix A^T does not satisfy these conditions.

30.7.1 The MART

The Kullback-Leibler, or KL distance [186] between positive numbers a and b is

$$KL(a,b) = a\log\frac{a}{b} + b - a. \tag{30.29}$$

We also define $KL(a,0) = +\infty$ and $KL(0,b) = b$. Extending to non-negative vectors $a = (a_1, ..., a_J)^T$ and $b = (b_1, ..., b_J)^T$, we have

$$KL(a,b) = \sum_{j=1}^{J} KL(a_j, b_j) = \sum_{j=1}^{J} \Big(a_j \log\frac{a_j}{b_j} + b_j - a_j\Big).$$

The MART is an iterative algorithm for finding a non-negative solution of the system $Px = y$, for an I by J matrix P with non-negative entries and vector y with positive entries. We also assume that

$$p_j = \sum_{i=1}^{I} P_{ij} > 0,$$

for all $i = 1, ..., I$. When discussing the MART, we say that the system $Px = y$ is *consistent* when it has non-negative solutions. We consider two different versions of the MART.

30.7.1.1 MART I

The iterative step of the first version of MART, which we shall call MART I, is the following: for $k = 0, 1, ...$, and $i = k(\mathrm{mod}\, I) + 1$, let

$$x_j^{k+1} = x_j^k \Big(\frac{y_i}{(Px^k)_i}\Big)^{P_{ij}/m_i},$$

for $j = 1, ..., J$, where the parameter m_i is defined to be

$$m_i = \max\{P_{ij} | j = 1, ..., J\}.$$

The MART I algorithm converges, in the consistent case, to the non-negative solution for which the KL distance $KL(x, x^0)$ is minimized.

30.7.1.2 MART II

The iterative step of the second version of MART, which we shall call MART II, is the following: for $k = 0, 1, ...$, and $i = k(\mathrm{mod}\, I) + 1$, let

$$x_j^{k+1} = x_j^k \Big(\frac{y_i}{(Px^k)_i} \Big)^{P_{ij}/p_j n_i},$$

for $j = 1, ..., J$, where the parameter n_i is defined to be

$$n_i = \max\{P_{ij} p_j^{-1} | j = 1, ..., J\}.$$

The MART II algorithm converges, in the consistent case, to the non-negative solution for which the KL distance

$$\sum_{j=1}^{J} p_j KL(x_j, x_j^0)$$

is minimized.

30.7.2 Using the MART to Solve the DGP Problem

Let the $(n+1)$ by m matrix A^T have the entries $A_{ji} = a_{ij}$, for $j = 1, ..., m$ and $i = 1, ..., n$, and $A_{(m+1),i} = 1$. Let u be the column vector with entries $u_j = 0$, for $j = 1, ..., m$, and $u_{m+1} = 1$.

The entries on the bottom row of A^T are all one, as is the bottom entry of the column vector u, since these entries correspond to the equation $\sum_{i=1}^{I} \delta_i = 1$. By adding suitably large positive multiples of this last equation to the other equations in the system, we obtain an equivalent system, $B^T \delta = s$, for which the new matrix B^T and the new vector s have only positive entries. Now we can apply the MART I algorithm to the system $B^T \delta = s$, letting $P = B^T$, $p_i = \sum_{j=1}^{J+1} B_{ij}$, $\delta = x$, $x^0 = c$ and $y = s$. In the consistent case, the MART I algorithm will find the non-negative solution that minimizes $KL(x, x^0)$, so we select $x^0 = c$. Then the MART I algorithm finds the non-negative δ^* satisfying $B^T \delta^* = s$, or, equivalently, $A^T \delta^* = u$, for which the KL distance

$$KL(\delta, c) = \sum_{i=1}^{I} \Big(\delta_i \log \frac{\delta_i}{c_i} + c_i - \delta_i \Big)$$

is minimized. Since we know that

$$\sum_{i=1}^{I} \delta_i = 1,$$

it follows that minimizing $KL(\delta, c)$ is equivalent to maximizing $v(\delta)$. Using δ^*, we find the optimal t^* solving the GP problem.

For example, the linear system of equations $A^T\delta = u$ corresponding to the posynomial in Equation (30.1) is

$$A^T\delta = u = \begin{bmatrix} -1 & 1 & 1 & 0 \\ -1 & 0 & 1 & 1 \\ -1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \\ \delta_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}.$$

Adding two times the last row to the other rows, the system becomes

$$B^T\delta = s = \begin{bmatrix} 1 & 3 & 3 & 2 \\ 1 & 2 & 3 & 3 \\ 1 & 3 & 2 & 3 \\ 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \\ \delta_4 \end{bmatrix} = \begin{bmatrix} 2 \\ 2 \\ 2 \\ 1 \end{bmatrix}.$$

The matrix B^T and the vector s are now positive. We are ready to apply the MART.

The MART iteration is as follows. With $j = k(\mathrm{mod}\,(J+1)) + 1$, $m_j = \max\{B_{ij}\,|i = 1, 2, ..., I\}$ and $k = 0, 1, ...$, let

$$\delta_i^{k+1} = \delta_i^k \Big(\frac{s_j}{(B^T\delta^k)_j}\Big)^{m_j^{-1}B_{ij}}.$$

The optimal δ^* is $\delta^* = (.4, .2, .2, .2)^T$, the optimal t^* is $t^* = (2, 1, .5)$, and the lowest cost is one hundred dollars.

30.8 Constrained Geometric Programming

Consider now the following variant of the problem of transporting the gravel across the river. Suppose that the bottom and the two sides will be constructed for free from scrap metal, but only four square yards are available. The cost function to be minimized becomes

$$g_0(t) = \frac{40}{t_1t_2t_3} + 40t_2t_3, \tag{30.30}$$

and the constraint is

$$g_1(t) = \frac{t_1t_3}{2} + \frac{t_1t_2}{4} \leq 1. \tag{30.31}$$

With $\delta_1 > 0$, $\delta_2 > 0$, and $\delta_1 + \delta_2 = 1$, we write

$$g_0(t) = \delta_1 \frac{40}{\delta_1 t_1 t_2 t_3} + \delta_2 \frac{40t_2t_3}{\delta_2}. \tag{30.32}$$

Since $0 \leq g_1(t) \leq 1$, we have

$$g_0(t) \geq \Big(\delta_1 \frac{40}{\delta_1 t_1 t_2 t_3} + \delta_2 \frac{40 t_2 t_3}{\delta_2}\Big)\Big(g_1(t)\Big)^{\lambda}, \tag{30.33}$$

for any positive λ. The GAGM Inequality then tells us that

$$g_0(t) \geq \left(\Big(\frac{40}{\delta_1 t_1 t_2 t_3}\Big)^{\delta_1}\Big(\frac{40 t_2 t_3}{\delta_2}\Big)^{\delta_2}\right)\Big(g_1(t)\Big)^{\lambda}, \tag{30.34}$$

so that

$$g_0(t) \geq \left(\Big(\frac{40}{\delta_1}\Big)^{\delta_1}\Big(\frac{40}{\delta_2}\Big)^{\delta_2}\right) t_1^{-\delta_1} t_2^{\delta_2-\delta_1} t_3^{\delta_2-\delta_1}\Big(g_1(t)\Big)^{\lambda}. \tag{30.35}$$

From the GAGM Inequality, we also know that, for $\delta_3 > 0$, $\delta_4 > 0$ and $\lambda = \delta_3 + \delta_4$,

$$\Big(g_1(t)\Big)^{\lambda} \geq (\lambda)^{\lambda}\left(\Big(\frac{1}{2\delta_3}\Big)^{\delta_3}\Big(\frac{1}{4\delta_4}\Big)^{\delta_4}\right) t_1^{\delta_3+\delta_4} t_2^{\delta_4} t_3^{\delta_3}. \tag{30.36}$$

Combining the inequalities in (30.35) and (30.36), we obtain

$$g_0(t) \geq v(\delta) t_1^{-\delta_1+\delta_3+\delta_4} t_2^{-\delta_1+\delta_2+\delta_4} t_3^{-\delta_1+\delta_2+\delta_3}, \tag{30.37}$$

with

$$v(\delta) = \Big(\frac{40}{\delta_1}\Big)^{\delta_1}\Big(\frac{40}{\delta_2}\Big)^{\delta_2}\Big(\frac{1}{2\delta_3}\Big)^{\delta_3}\Big(\frac{1}{4\delta_4}\Big)^{\delta_4}\Big(\delta_3+\delta_4\Big)^{\delta_3+\delta_4}, \tag{30.38}$$

and $\delta = (\delta_1, \delta_2, \delta_3, \delta_4)$. If we can find a positive vector δ with

$$\begin{aligned} \delta_1 + \delta_2 &= 1, \\ \delta_3 + \delta_4 &= \lambda, \\ -\delta_1 + \delta_3 + \delta_4 &= 0, \\ -\delta_1 + \delta_2 + \delta_4 &= 0 \\ -\delta_1 + \delta_2 + \delta_3 &= 0, \end{aligned} \tag{30.39}$$

then

$$g_0(t) \geq v(\delta). \tag{30.40}$$

In this particular case, there is a unique positive δ satisfying the equations (30.39), namely

$$\delta_1^* = \frac{2}{3}, \delta_2^* = \frac{1}{3}, \delta_3^* = \frac{1}{3}, \text{and } \delta_4^* = \frac{1}{3}, \tag{30.41}$$

and

$$v(\delta^*) = 60. \tag{30.42}$$

Therefore, $g_0(t)$ is bounded below by 60. If there is t^* such that

$$g_0(t^*) = 60, \tag{30.43}$$

then we must have

$$g_1(t^*) = 1, \tag{30.44}$$

and equality in the GAGM Inequality. Consequently,

$$\frac{3}{2}\frac{40}{t_1^* t_2^* t_3^*} = 3(40 t_2^* t_3^*) = 60, \tag{30.45}$$

and

$$\frac{3}{2} t_1^* t_3^* = \frac{3}{4} t_1^* t_2^* = K. \tag{30.46}$$

Since $g_1(t^*) = 1$, we must have $K = \frac{3}{2}$. We solve these equations by taking logarithms, to obtain the solution

$$t_1^* = 2,\ t_2^* = 1, \text{ and } t_3^* = \frac{1}{2}. \tag{30.47}$$

The change of variables $t_j = e^{x_j}$ converts the constrained (GP) problem into a constrained convex programming problem. The theory of the constrained (GP) problem can then be obtained as a consequence of the theory for the convex programming problem.

30.9 Exercises

Ex. 30.1 *Show that there is no solution to the problem of minimizing the function*

$$g(t_1, t_2) = \frac{2}{t_1 t_2} + t_1 t_2 + t_1, \tag{30.48}$$

over $t_1 > 0$, $t_2 > 0$.

Ex. 30.2 *Minimize the function*

$$g(t_1, t_2) = \frac{1}{t_1 t_2} + t_1 t_2 + t_1 + t_2, \tag{30.49}$$

over $t_1 > 0$, $t_2 > 0$. *This will require some iterative numerical method for solving equations.*

Ex. 30.3 *Program the MART algorithm and use it to verify the assertions made previously concerning the solutions of the two numerical examples.*

Chapter 31

Appendix: Fourier Transforms and the FFT

31.1 Chapter Summary

The *Fourier transform* of a complex-valued function $f(x)$ of the real variable x is defined as

$$F(\omega) = \int_{-\infty}^{\infty} f(x) e^{ix\omega} dx. \tag{31.1}$$

If we have $F(\omega)$, we can obtain $f(x)$ again via the *Fourier Inversion Formula*,

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(\omega) e^{-ix\omega} d\omega. \tag{31.2}$$

In many applications, particularly in remote sensing, what we are able to measure are values of $f(x)$, and what we really want is the function $F(\omega)$. This is the case in medical tomography, magnetic-resonance imaging, sonar, radar, optical and radio astronomy, and many other areas. Because our measurements are finite in number, the problem becomes how to estimate $F(\omega)$ from finitely many values of $f(x)$. The *fast Fourier transform* (FFT) is a fast algorithm for calculating one such estimate, the *discrete Fourier transform*. Discovered in 1965 by Cooley and Tukey, the FFT has revolutionized signal and image processing. The man in Figure 31.1 is John Tukey.

31.2 Non-periodic Convolution

Recall the algebra problem of multiplying one polynomial by another. Suppose

$$A(x) = a_0 + a_1 x + ... + a_M x^M$$

and

$$B(x) = b_0 + b_1 x + ... + b_N x^N.$$

Let $C(x) = A(x)B(x)$. With

$$C(x) = c_0 + c_1 x + ... + c_{M+N} x^{M+N},$$

each of the coefficients c_j, $j = 0, ..., M+N$, can be expressed in terms of the a_m and b_n (an easy exercise!). The vector $c = (c_0, ..., c_{M+N})$ is called the *non-periodic convolution* of the vectors $a = (a_0, ..., a_M)$ and $b = (b_0, ..., b_N)$. Non-periodic convolution can be viewed as a particular case of periodic convolution, as we shall see.

31.3 The DFT as a Polynomial

Given the complex numbers $f_0, f_1, ..., f_{N-1}$, which may or may not be measured values of $f(x)$, we form the vector $\mathbf{f} = (f_0, f_1, ..., f_{N-1})^T$. The DFT of the vector $\mathbf{f}$ is the function

$$DFT_{\mathbf{f}}(\omega) = \sum_{n=0}^{N-1} f_n e^{in\omega},$$

defined for ω in the interval $[0, 2\pi)$. Because $e^{in\omega} = (e^{i\omega})^n$, we can write the DFT as a polynomial

$$DFT_{\mathbf{f}}(\omega) = \sum_{n=0}^{N-1} f_n (e^{i\omega})^n.$$

If we have a second vector, say $\mathbf{d} = (d_0, d_1, ..., d_{N-1})^T$, then we define $DFT_{\mathbf{d}}(\omega)$ similarly. When we multiply $DFT_{\mathbf{f}}(\omega)$ by $DFT_{\mathbf{d}}(\omega)$, we are multiplying two polynomials together, so the result is a sum of powers of the form

$$c_0 + c_1 e^{i\omega} + c_2 (e^{i\omega})^2 + ... + c_{2N-2} (e^{i\omega})^{2N-2}, \tag{31.3}$$

for

$$c_j = f_0 d_j + f_1 d_{j-1} + ... + f_j d_0.$$

This is *non-periodic convolution* again. In the next section, we consider what happens when, instead of using arbitrary values of ω, we consider only the N special values $\omega_k = \frac{2\pi}{N}k$, $k = 0, 1, ..., N-1$. Because of the periodicity of the complex exponential function, we have

$$(e^{i\omega_k})^{N+j} = (e^{i\omega_k})^j,$$

for each k. As a result, all the powers higher than $N-1$ that showed up in the previous multiplication in Equation (31.3) now become equal to lower powers, and the product now only has N terms, instead of the $2N-1$ terms we got previously. When we calculate the coefficients of these powers, we find that we get more than we got when we did the non-periodic convolution. Now what we get is called *periodic convolution*.

31.4 The Vector DFT and Periodic Convolution

As we just discussed, non-periodic convolution is another way of looking at the multiplication of two polynomials. This relationship between convolution on the one hand and multiplication on the other is a fundamental aspect of convolution. Whenever we have a convolution we should ask what related mathematical objects are being multiplied. We ask this question now with regard to periodic convolution; the answer turns out to be the *vector discrete Fourier transform* (vDFT).

31.4.1 The Vector DFT

Let $\mathbf{f} = (f_0, f_1, ..., f_{N-1})^T$ be a column vector whose entries are N arbitrary complex numbers. For $k = 0, 1, ..., N-1$, we let

$$F_k = \sum_{n=0}^{N-1} f_n e^{2\pi ikn/N} = DFT_{\mathbf{f}}(\omega_k). \tag{31.4}$$

Then we let $\mathbf{F} = (F_0, F_1, ..., F_{N-1})^T$ be the column vector with the N complex entries F_k. The vector $\mathbf{F}$ is called the *vector discrete Fourier transform* of the vector $\mathbf{f}$, and we denote it by $\mathbf{F} = vDFT_{\mathbf{f}}$.

As we can see from Equation (31.4), there are N multiplications involved in the calculation of each F_k, and there are N values of k, so it would seem that, in order to calculate the vector DFT of $\mathbf{f}$, we need N^2 multiplications.

In many applications, N is quite large and calculating the vector $\mathbf{F}$ using the definition would be unrealistically time-consuming. The *fast Fourier transform* algorithm (FFT), to be discussed later, gives a quick way to calculate the vector $\mathbf{F}$ from the vector $\mathbf{f}$. The FFT, usually credited to Cooley and Tukey, was discovered in the mid-1960's and revolutionized signal and image processing.

31.4.2 Periodic Convolution

Given the N by 1 vectors $\mathbf{f}$ and $\mathbf{d}$ with complex entries f_n and d_n, respectively, we define a third N by 1 vector $\mathbf{f} * \mathbf{d}$, the *periodic convolution* of $\mathbf{f}$ and $\mathbf{d}$, to have the entries

$$(\mathbf{f} * \mathbf{d})_n = f_0 d_n + f_1 d_{n-1} + ... + f_n d_0 + f_{n+1} d_{N-1} + ... + f_{N-1} d_{n+1}, \quad (31.5)$$

for $n = 0, 1, ..., N-1$.

Notice that the term on the right side of Equation (31.5) is the sum of all products of entries, one from $\mathbf{f}$ and one from $\mathbf{d}$, where the sum of their respective indices is either n or $n+N$.

In the exercises that follow we investigate properties of the vector DFT and relate it to periodic convolution. It is not an exaggeration to say that these two exercises are the most important ones in signal processing.

Ex. 31.1 *Let* $\mathbf{F} = vDFT_{\mathbf{f}}$ *and* $\mathbf{D} = vDFT_{\mathbf{d}}$. *Define a third vector* $\mathbf{E}$ *having for its kth entry* $E_k = F_k D_k$, *for* $k = 0, ..., N-1$. *Show that* $\mathbf{E}$ *is the vDFT of the vector* $\mathbf{f} * \mathbf{d}$.

The vector $vDFT_{\mathbf{f}}$ can be obtained from the vector $\mathbf{f}$ by means of matrix multiplication by a certain matrix G, called the *DFT matrix*. The matrix G has an inverse that is easily computed and can be used to go from $\mathbf{F} = vDFT_{\mathbf{f}}$ back to the original $\mathbf{f}$. The details are in Exercise 31.2.

Ex. 31.2 *Let* G *be the* N *by* N *matrix whose entries are* $G_{jk} = e^{i(j-1)(k-1)2\pi/N}$. *The matrix* G *is sometimes called the DFT matrix. Show that the inverse of* G *is* $G^{-1} = \frac{1}{N}G^{\dagger}$, *where* $G^{\dagger}$ *is the conjugate transpose of the matrix* G. *Then* $\mathbf{f} * \mathbf{d} = G^{-1}\mathbf{E} = \frac{1}{N}G^{\dagger}\mathbf{E}$.

As mentioned previously, non-periodic convolution is really a special case of periodic convolution. Extend the $M+1$ by 1 vector a to an $M+N+1$ by 1 vector by appending N zero entries; similarly, extend the vector b to an $M+N+1$ by 1 vector by appending zeros. The vector c is now the periodic convolution of these extended vectors. Therefore, since we have an efficient algorithm for performing periodic convolution, namely the Fast Fourier Transform algorithm (FFT), we have a fast way to do the periodic (and thereby non-periodic) convolution and polynomial multiplication.

31.5 The Fast Fourier Transform (FFT)

A fundamental problem in signal processing is to estimate the function $F(\omega)$ from finitely many values of its (inverse) Fourier transform, $f(x)$. As we have seen, the DFT is one such estimate. The *fast Fourier transform* (FFT), discovered in 1965 by Cooley and Tukey, is an important and efficient algorithm for calculating the vector DFT [101]. John Tukey has been quoted as saying that his main contribution to this discovery was the firm and often voiced belief that such an algorithm must exist.

To illustrate the main idea underlying the FFT, consider the problem of evaluating a real polynomial $P(x)$ at a point, say $x = c$. Let the polynomial be

$$P(x) = a_0 + a_1 x + a_2 x^2 + ... + a_{2K} x^{2K},$$

where a_{2K} might be zero. Performing the evaluation efficiently by Horner's method,

$$P(c) = (((a_{2K} c + a_{2K-1})c + a_{2K-2})c + a_{2K-3})c + ...,$$

requires $2K$ multiplications, so the complexity is on the order of the degree of the polynomial being evaluated. But suppose we also want $P(-c)$. We can write

$$P(x) = (a_0 + a_2 x^2 + ... + a_{2K} x^{2K}) + x(a_1 + a_3 x^2 + ... + a_{2K-1} x^{2K-2})$$

or

$$P(x) = Q(x^2) + xR(x^2).$$

Therefore, we have $P(c) = Q(c^2) + cR(c^2)$ and $P(-c) = Q(c^2) - cR(c^2)$. If we evaluate $P(c)$ by evaluating $Q(c^2)$ and $R(c^2)$ separately, one more multiplication gives us $P(-c)$ as well. The FFT is based on repeated use of this idea, which turns out to be more powerful when we are using complex exponentials, because of their periodicity.

Say the data are $\{f_n = f(n), n = 0, ..., N-1\}$. The DFT estimate of $F(\omega)$ is the function $DFT_{\mathbf{f}}(\omega)$, defined for ω in $[0, 2\pi]$, and given by

$$DFT_{\mathbf{f}}(\omega) = \sum_{n=0}^{N-1} f(n) e^{in\omega}.$$

The DFT estimate $DFT(\omega)$ is data consistent; its inverse Fourier-transform value at $x = n$ is $f(n)$ for $n = 0, ..., N-1$. The DFT is also used in a more general context in which the f_n are not necessarily values of a function $f(x)$.

Given any complex N-dimensional column vector $\mathbf{f} = (f_0, f_1, ..., f_{N-1})^T$,

define the DFT of the vector $\mathbf{f}$ to be the function $DFT_{\mathbf{f}}(\omega)$, defined for ω in $[0, 2\pi)$, given by

$$DFT_{\mathbf{f}}(\omega) = \sum_{n=0}^{N-1} f_n e^{in\omega}.$$

Let $\mathbf{F}$ be the complex N-dimensional vector $\mathbf{F} = (F_0, F_1, ..., F_{N-1})^T$, where $F_k = DFT_{\mathbf{f}}(2\pi k/N), k = 0, 1, ..., N-1$. So the vector $\mathbf{F}$ consists of N values of the function $DFT_{\mathbf{f}}$, taken at N equispaced points $2\pi/N$ apart in $[0, 2\pi)$.

From the formula for $DFT_{\mathbf{f}}$ we have, for $k = 0, 1, ..., N-1$,

$$F_k = F(2\pi k/N) = \sum_{n=0}^{N-1} f_n e^{2\pi ink/N}. \tag{31.6}$$

To calculate a single F_k requires N multiplications; it would seem that to calculate all N of them would require N^2 multiplications. However, using the FFT algorithm, we can calculate vector $\mathbf{F}$ in approximately $N \log_2(N)$ multiplications.

Suppose that $N = 2M$ is even. We can rewrite Equation (31.6) as follows:

$$F_k = \sum_{m=0}^{M-1} f_{2m} e^{2\pi i(2m)k/N} + \sum_{m=0}^{M-1} f_{2m+1} e^{2\pi i(2m+1)k/N},$$

or, equivalently,

$$F_k = \sum_{m=0}^{M-1} f_{2m} e^{2\pi imk/M} + e^{2\pi ik/N} \sum_{m=0}^{M-1} f_{2m+1} e^{2\pi imk/M}. \tag{31.7}$$

Note that if $0 \leq k \leq M-1$ then

$$F_{k+M} = \sum_{m=0}^{M-1} f_{2m} e^{2\pi imk/M} - e^{2\pi ik/N} \sum_{m=0}^{M-1} f_{2m+1} e^{2\pi imk/M}, \tag{31.8}$$

so there is no additional computational cost in calculating the second half of the entries of $\mathbf{F}$, once we have calculated the first half. The FFT is the algorithm that results when we take full advantage of the savings obtainable by splitting a DFT calculating into two similar calculations of half the size.

We assume now that $N = 2^L$. Notice that if we use Equations (31.7) and (31.8) to calculate vector $\mathbf{F}$, the problem reduces to the calculation of two similar DFT evaluations, both involving half as many entries, followed by one multiplication for each of the k between 0 and $M-1$. We can split these in half as well. The FFT algorithm involves repeated splitting of the calculations of DFTs at each step into two similar DFTs, but with half the

number of entries, followed by as many multiplications as there are entries in either one of these smaller DFTs. We use recursion to calculate the cost $C(N)$ of computing $\mathbf{F}$ using this FFT method. From Equation (31.7) we see that $C(N) = 2C(N/2) + (N/2)$. Applying the same reasoning to get $C(N/2) = 2C(N/4) + (N/4)$, we obtain

$$C(N) = 2C(N/2) + (N/2) = 4C(N/4) + 2(N/2) = ...$$

$$= 2^L C(N/2^L) + L(N/2) = N + L(N/2).$$

Therefore, the cost required to calculate $\mathbf{F}$ is approximately $N \log_2 N$.

From our earlier discussion of discrete linear filters and convolution, we see that the FFT can be used to calculate the periodic convolution (or even the nonperiodic convolution) of finite length vectors.

Finally, let's return to the original context of estimating the Fourier transform $F(\omega)$ of function $f(x)$ from finitely many samples of $f(x)$. If we have N equispaced samples, we can use them to form the vector $\mathbf{f}$ and perform the FFT algorithm to get vector $\mathbf{F}$ consisting of N values of the DFT estimate of $F(\omega)$. It may happen that we wish to calculate more than N values of the DFT estimate, perhaps to produce a smooth looking graph. We can still use the FFT, but we must trick it into thinking we have more data that the N samples we really have. We do this by *zero-padding*. Instead of creating the N-dimensional vector $\mathbf{f}$, we make a longer vector by appending, say, J zeros to the data, to make a vector that has dimension $N + J$. The DFT estimate is still the same function of ω, since we have only included new zero coefficients as fake data; but, the FFT thinks we have $N + J$ data values, so it returns $N + J$ values of the DFT, at $N + J$ equispaced values of ω in $[0, 2\pi)$.

FIGURE 31.1: John Tukey: co-inventor of the FFT.

Chapter 32

Appendix: Hermitian and Normal Linear Operators

32.1 Chapter Summary

We saw previously that if the finite-dimensional vector space V has an orthonormal basis of eigenvectors of the linear operator T, then T is a normal operator. We need to prove the converse: if T is normal, then V has an orthonormal basis consisting of eigenvectors of T. Earlier, we proved this result using matrix representations of linear operators and Schur's Lemma. Now we give a proof within the context of linear operators themselves.

32.2 The Diagonalization Theorem

In this chapter we present a proof of the following theorem.

Theorem 32.1 *For a linear operator T on a finite-dimensional complex inner-product space V there is an orthonormal basis of eigenvectors if and only if T is normal.*

32.3 Invariant Subspaces

A subspace W of V is said to be *T-invariant* if Tw is in W whenever w is in W. For any T-invariant subspace W, the restriction of T to W, denoted T_W, is a linear operator on W.

For any subspace W, the *orthogonal complement* of W is the space $W^{\perp} = \{v | \langle w, v\rangle = 0, \text{for all } w \in W\}$.

Proposition 32.1 *Let W be a T-invariant subspace of V. Then*

- *(a) if T is self-adjoint, so is T_W;*
- *(b) $W^{\perp}$ is T^*-invariant;*
- *(c) if W is both T- and T^*-invariant, then $(T_W)^* = (T^*)_W$;*
- *(d) if W is both T- and T^*-invariant, and T is normal, then T_W is normal.*
- *(e) if T is normal and $Tx = \lambda x$, then $T^*x = \overline{\lambda}x$.*

Ex. 32.1 *Prove Proposition (32.1).*

Proposition 32.2 *If T is normal, $Tu^1 = \lambda_1 u^1$, $Tu^2 = \lambda_2 u^2$, and $\lambda_1 \neq \lambda_2$, then $\langle u^1, u^2\rangle = 0$.*

Ex. 32.2 *Prove Proposition 32.2. Hint: use (e) of Proposition 32.1.*

32.4 Proof of the Diagonalization Theorem

We turn now to the proof of the theorem.

Proof of Theorem 32.1 The proof is by induction on the dimension of the inner-product space V. To begin with, let $N = 1$, so that V is simply the

span of some unit vector x. Then any linear operator T on V has $Tx = \lambda x$, for some λ, and the set $\{x\}$ is an orthonormal basis for V.

Now suppose that the theorem is true for every inner-product space of dimension $N-1$. We know that every linear operator T on V has at least one eigenvector, say x^1, since its characteristic polynomial has at least one distinct root λ_1 in C. Take x^1 to be a unit vector. Let W be the span of the vector x^1, and $W^\perp$ the orthogonal complement of W. Since $Tx^1 = \lambda_1 x^1$ and T is normal, we know that $T^* x^1 = \overline{\lambda_1} x^1$. Therefore, both W and $W^\perp$ are T- and T^*-invariant. Therefore, $T_{W^\perp}$ is normal on $W^\perp$. By the induction hypothesis, we know that $W^\perp$ has an orthonormal basis consisting of $N-1$ eigenvectors of $T_{W^\perp}$, and, therefore, of T. Augmenting this set with the original x^1, we get an orthonormal basis for all of V. ∎

32.5 Corollaries

The theorem has several important corollaries.

Corollary 32.1 *A self-adjoint linear operator T on a finite-dimensional complex inner-product space V has an orthonormal basis of eigenvectors.*

Corollary 32.2 *Let T be a linear operator on a finite-dimensional real inner-product space V. Then V has an orthonormal basis consisting of eigenvectors of T if and only if T is self-adjoint.*

Corollary 32.3 *Let A be a normal matrix. Then there is a unitary matrix U and diagonal matrix L such that $A = ULU^\dagger$.*

Proving the existence of the orthonormal basis uses essentially the same argument as the induction proof given earlier. The eigenvalues of a self-adjoint linear operator T on a finite-dimensional complex inner-product space are real numbers. If T be a linear operator on a finite-dimensional real inner-product space V and V has an orthonormal basis $\mathcal{U} = \{u^1, ..., u^N\}$ consisting of eigenvectors of T, then we have

$$Tu^n = \lambda_n u^n = \overline{\lambda_n} u^n = T^* u^n,$$

so, since $T = T^*$ on each member of the basis, these operators are the same everywhere, so $T = T^*$ and T is self-adjoint.

Definition 32.1 *A linear operator P on a finite-dimensional inner-product space is a* perpendicular projection *if*

$$P^2 = P = P^*.$$

Corollary 32.4 (The Spectral Theorem) *Let T be a normal operator on a finite-dimensional inner-product space. Then T can be written as*

$$T = \sum_{m=1}^{M} \lambda_m P_m, \tag{32.1}$$

where λ_m, $m = 1, ..., M$ are the distinct eigenvalues of T, P_m is the perpendicular projection

$$P_m = \sum_{n \in I_m} u^n (u^n)^\dagger, \tag{32.2}$$

and

$$I_m = \{n | \lambda_n = \lambda_m\}.$$

Corollary 32.5 *Let T be a normal operator on a finite-dimensional inner-product space. Then there is a complex polynomial $f(z)$ such that*

$$T^* = f(T).$$

Proof: Let $f(z)$ be any polynomial such that $f(\lambda_m) = \overline{\lambda_m}$, for each $m = 1, ..., M$. The assertion then follows, since

$$T^* = \sum_{m=1}^{M} \overline{\lambda_m} P_m,$$

and $P_m P_k = 0$, for $m \neq k$. ∎

32.6 A Counter-Example

We present now an example of a real 2 by 2 matrix A with $A^T A = A A^T$, but with no eigenvectors in R^2. Take $0 < \theta < \pi$ and A to be the matrix

$$A = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}. \tag{32.3}$$

This matrix represents rotation through an angle of θ in R^2. Its transpose represents rotation through the angle $-\theta$. These operations obviously can be done in either order, so the matrix A is normal. But there is no non-zero vector in R^2 that is an eigenvector. Clearly, A is not symmetric; but it is skew-symmetric.

32.7 Simultaneous Diagonalization

Any linear operator T on a finite-dimensional inner-product space can be written as $T = R + iS$, where both R and S are Hermitian linear operators; simply take $R = \frac{1}{2}(T + T^*)$ and $S = \frac{1}{2i}(T - T^*)$.

Ex. 32.3 *Show that T is a normal operator if and only if $RS = SR$.*

Theorem 32.2 *Let T and U be commuting normal linear operators on a finite-dimensional inner-product space V. There there is an orthonormal basis for V consisting of vectors that are simultaneously eigenvectors for T and for U.*

Proof: For each m let W_m be the range of the perpendicular projection P_m in the spectral theorem expansion for T; that is,

$$W_m = \{x \in V | Tx = \lambda_m x\}.$$

It is easy to see that, for each x in W_m, the vector Ux is in W_m; therefore, the sets W_m are T- and U-invariant. It follows along the lines of our proof of the spectral theorem that the restriction of U to each of the subspaces W_m is a normal operator. Therefore, each W_m has an orthonormal basis consisting of eigenvectors of U. Combining these bases for the W_m gives the desired basis for V. ∎

When T is normal, we have $RS = SR$, so there is an orthonormal basis for V consisting of simultaneous eigenvectors for R and S. It follows that these basis vectors are eigenvectors for T as well. This shows that the spectral theorem for normal operators can be derived from the spectral theorem for Hermitian operators, once we have the simultaneous-diagonalization theorem for commuting Hermitian operators.

It can be shown that, for any family of commuting normal operators on V, there is an orthonormal basis of simultaneous eigenvectors. The recent article by Bouten, van Handel and James [25] describes the use of this result in quantum filtering.

32.8 Quadratic Forms and Congruent Operators

If Q is a Hermitian positive-definite N by N matrix, then the function

$$\phi(x, y) = y^\dagger Q x = \langle x, y \rangle_Q$$

is an inner product on $\mathbb{C}^N$, and the quadratic form

$$\hat{\phi}(x) = x^\dagger Q x = \langle x, x \rangle_Q = \|x\|_Q^2,$$

is the square of the Q-norm. If S is an arbitrary N by N matrix, then the function $\hat{\phi}(x) = x^\dagger S x$ will not be a norm, generally, and $\phi(x, y) = y^\dagger S x$ will not be an inner product, unless S is Hermitian and positive-definite. However, the function $\phi(x, y) = y^\dagger S x$ will still possess some of the properties of an inner product. Such functions are called *sesquilinear forms* or *sesquilinear functionals*.

32.8.1 Sesquilinear Forms

Let V be any complex vector space. A *sesquilinear functional* $\phi(x, y)$ of two variables in V is linear in the first variable and conjugate-linear in the second; that is,

$$\phi(x, \alpha_1 y^1 + \alpha_2 y^2) = \overline{\alpha_1}\phi(x, y^1) + \overline{\alpha_2}\phi(x, y^2);$$

the term *sesquilinear* means *one and one-half linear*. An inner product on V is a special kind of sesquilinear functional.

32.8.2 Quadratic Forms

Any sesquilinear functional has an associated *quadratic form* given by

$$\hat{\phi}(x) = \phi(x, x).$$

If P is any invertible linear operator on V, we can define a new quadratic form by

$$\hat{\phi}_P(x) = \phi(Px, Px).$$

32.8.3 Congruent Linear Operators

Let T be a linear operator on an inner product space V. Then T can be used to define a sesquilinear functional $\phi_T(x, y)$ according to

$$\phi_T(x, y) = \langle Tx, y \rangle. \tag{32.4}$$

For this sesquilinear functional $\phi_T(x, y)$, we have

$$(\hat{\phi}_T)_P(x) = \phi_T(Px, Px) = \langle TPx, Px \rangle = \langle P^*TPx, x \rangle.$$

We say that a linear operator U on V is *congruent* to T if there is an invertible linear operator P with $U = P^*TP$.

In order for the sesquilinear functional $\phi_T(x, y) = \langle Tx, y\rangle$ to be an inner product, it is necessary and sufficient that T be positive-definite; that is, for all x in V,

$$\phi_T(x, x) = \langle Tx, x\rangle \geq 0,$$

with equality if and only if $x = 0$.

32.8.4 Congruent Matrices

Now let $V = \mathbb{C}^N$, with the usual basis and inner product. Linear operators T, U and P are identified with their corresponding matrix representations. We then say that the matrix B is *congruent* to matrix A if there is an invertible matrix P for which $B = P^\dagger AP$.

32.8.5 Does ϕ_T Determine T?

Let T and U be linear operators on an inner product space V. Is it possible for

$$\langle Tx, x\rangle = \langle Ux, x\rangle,$$

for all x in the inner product space V, and yet have $T \neq U$? As we shall see, the answer is "No". First, we answer a simpler question. Is it possible for

$$\langle Tx, y\rangle = \langle Ux, y\rangle,$$

for all x and y, with $T \neq U$? The answer again is "No".

Ex. 32.4 *Show that*

$$\langle Tx, y\rangle = \langle Ux, y\rangle,$$

for all x and y, implies that $T = U$.

We can use the result of the exercise to answer our first question, but first, we need the *polarization identity*.

Ex. 32.5 *Establish the* polarization identity*:*

$$\langle Tx, y\rangle = \frac{1}{4}\langle T(x + y), x + y\rangle - \frac{1}{4}\langle T(x - y), x - y\rangle$$

$$+\frac{i}{4}\langle T(x + iy), x + iy\rangle - \frac{i}{4}\langle T(x - iy), x - iy\rangle.$$

Ex. 32.6 *Show that the answer to our first question is "No"; the quadratic form determines the operator.*

32.8.6 A New Sesquilinear Functional

Given any sesquilinear functional $\phi(x, y)$ and two linear operators P and Q on V, we can define a second sesquilinear functional

$$\psi(x, y) = \phi(Px, Qy).$$

For the sesquilinear functional ϕ_T, we have

$$\psi(x, y) = \phi_T(Px, Qy) = \langle TPx, Qy \rangle = \langle Q^*TPx, y \rangle.$$

Chapter 33

Appendix: Sturm-Liouville Problems

33.1 Chapter Summary

Previously, we discussed self-adjoint linear operators on an inner product space. An important application of this theory is the analysis of linear ordinary differential equations in Sturm-Liouville form. Now the linear operators involved are differential operators, the members of the inner product space are twice differentiable functions of a single variable, and the inner product is defined in terms of an integration. The eigenvectors of the differential operators are *eigenfunctions*. The expansion of members of the inner product space in terms of bases of eigenvectors becomes the famous expansion of functions as sums of Bessel functions, Legendre polynomials and so on.

33.2 Second-Order Linear ODE

The most general form of the second-order linear homogeneous ordinary differential equation with variable coefficients is

$$R(x)y''(x) + P(x)y'(x) + Q(x)y(x) = 0. \tag{33.1}$$

Many differential equations of this type arise when we employ the technique of separating the variables to solve a partial differential equation.

33.2.1 The Standard Form

Of course, dividing through by the function $R(x)$ and renaming the coefficient functions, we can also write Equation (33.1) in the *standard* form as

$$y''(x) + P(x)y'(x) + Q(x)y(x) = 0. \tag{33.2}$$

There are other equivalent forms of Equation (33.1).

33.2.2 The Sturm-Liouville Form

Let $S(x) = \exp(-F(x))$, where $F'(x) = (R'(x) - P(x))/R(x)$. Then we have

$$\frac{d}{dx}(S(x)R(x)) = S(x)P(x).$$

From Equation (33.1) we obtain

$$S(x)R(x)y''(x) + S(x)P(x)y'(x) + S(x)Q(x)y(x) = 0,$$

so that

$$\frac{d}{dx}(S(x)R(x)y'(x)) + S(x)Q(x)y(x) = 0,$$

which then has the form

$$\frac{d}{dx}(p(x)y'(x)) - w(x)q(x)y(x) + \lambda w(x)y(x) = 0, \tag{33.3}$$

where $w(x) > 0$ and λ is a constant. Rewriting Equation (33.3) as

$$-\frac{1}{w(x)}\Big(\frac{d}{dx}(p(x)y'(x))\Big) + q(x)y(x) = \lambda y(x), \tag{33.4}$$

suggests an analogy with the linear algebra eigenvalue problem

$$Ax = \lambda x, \tag{33.5}$$

where A is a square matrix, λ is an eigenvalue of A, and $x \neq 0$ is an associated eigenvector. It also suggests that we study the linear differential operator

$$(Ly)(x) = -\frac{1}{w(x)}\Big(\frac{d}{dx}(p(x)y'(x))\Big) + q(x)y(x) \tag{33.6}$$

to see if we can carry the analogy with linear algebra further.

33.3 Inner Products and Self-Adjoint Differential Operators

For the moment, let V_0 be the vector space of complex-valued integrable functions $f(x)$, defined for $a \leq x \leq b$, for which

$$\int_a^b |f(x)|^2 dx < \infty.$$

For any f and g in V_0 the inner product of f and g is then

$$\langle f, g\rangle = \int_a^b f(x)\overline{g(x)}dx.$$

The linear differential operator

$$Ty = y''$$

is defined for the subspace V_1 of functions $y(x)$ in V_0 that are twice continuously differentiable. Now let V be the subspace of V_1 consisting of all $y(x)$ with $y(a) = y(b) = 0$.

Proposition 33.1 *The operator $Ty = y''$ is self-adjoint on V.*

Proof: We need to show that

$$\langle Ty, z\rangle = \int_a^b y''(x)z(x)dx = \int_a^b y(x)z''(x)dx = \langle y, Tz\rangle,$$

for all $y(x)$ and $z(x)$ in V. This follows immediately from two applications of integration by parts and the restrictions $y(a) = z(a) = y(b) = z(b) = 0$. ∎

It is useful to note that

$$\langle Ty, y\rangle = -\int_a^b |y'(x)|^2 dx \leq 0,$$

for all $y(x)$ in V, which prompts us to say that the differential operator $(-T)y = -y''$ is *non-negative definite*. We then expect all eigenvalues of $-T$ to be non-negative. We know, in particular, that solutions of

$$-y''(x) = \lambda y(x),$$

with $y(0) = y(1) = 0$ are $y_m(x) = \sin(m\pi x)$, and the eigenvalues are $\lambda_m = m^2\pi^2$.

We turn now to the differential operator L given by Equation (33.6). We take V_0 to be all complex-valued integrable functions $f(x)$ with

$$\int_a^b |f(x)|^2 w(x)dx < \infty.$$

We let the inner product of any $f(x)$ and $g(x)$ in V_0 be

$$\langle f, g\rangle = \int_a^b f(x)\overline{g(x)}w(x)dx. \tag{33.7}$$

Let V_1 be all functions in V_0 that are twice continuously differentiable, and V all the functions $y(x)$ in V_1 with $y(a) = y(b) = 0$. We then have the following result.

Theorem 33.1 *The operator L given by Equation (33.6) is self-adjoint on the inner product space V.*

Proof: It is easily seen that

$$(Ly)z - y(Lz) = \frac{1}{w(x)}\frac{d}{dx}(pyz' - py'z).$$

Therefore,

$$\int_a^b \Big((Ly)z - y(Lz)\Big)w(x)dx = (pyz' - py'z)|_a^b = 0.$$

Therefore, $L^* = L$ on V. ∎

It is interesting to note that

$$\langle Ly, y\rangle = \int_a^b p(y')^2 dx + \int_a^b qy^2 dx,$$

so that, if we have $p(x) \geq 0$ and $q(x) \geq 0$, then the operator L is non-negative-definite and we expect all its eigenvalues to be non-negative.

33.4 Orthogonality

Once again, let V be the space of all twice continuously differentiable functions $y(x)$ on $[a, b]$ with $y(a) = y(b) = 0$. Let λ_m and λ_n be distinct eigenvalues of the linear differential operator L given by Equation (33.6), with associated eigenfunctions $u_m(x)$ and $u_n(x)$, respectively. Let the inner product on V be given by Equation (33.7).

Theorem 33.2 *The eigenfunctions $u_m(x)$ and $u_n(x)$ are orthogonal.*

Proof: We have

$$\frac{d}{dx}(p(x)u'_m(x)) - w(x)q(x)u_m(x) = -\lambda_m u_m(x)w(x),$$

and

$$\frac{d}{dx}(p(x)u'_n(x)) - w(x)q(x)u_n(x) = -\lambda_m u_n(x)w(x),$$

so that

$$u_n(x)\frac{d}{dx}(p(x)u'_m(x)) - w(x)q(x)u_m(x)u_n(x) = -\lambda_m u_m(x)u_n(x)w(x)$$

and

$$u_m(x)\frac{d}{dx}(p(x)u'_n(x)) - w(x)q(x)u_m(x)u_n(x) = -\lambda_n u_m(x)u_n(x)w(x).$$

Subtracting one equation from the other, we get

$$u_n(x)\frac{d}{dx}(p(x)u'_m(x)) - u_m(x)\frac{d}{dx}(p(x)u'_n(x)) = (\lambda_n - \lambda_m)u_m(x)u_n(x)w(x).$$

The left side of the previous equation can be written as

$$u_n(x)\frac{d}{dx}(p(x)u'_m(x)) - u_m(x)\frac{d}{dx}(p(x)u'_n(x))$$

$$= \frac{d}{dx}\Big(p(x)u_n(x)u'_m(x) - p(x)u_m(x)u'_n(x)\Big).$$

Therefore,

$$(\lambda_n - \lambda_m)\int_a^b u_m(x)u_n(x)w(x)dx =$$

$$\Big(p(x)u_n(x)u'_m(x) - p(x)u_m(x)u'_n(x)\Big)\Big|_a^b = 0. \tag{33.8}$$

Since $\lambda_m \neq \lambda_n$, it follows that

$$\int_a^b u_m(x)u_n(x)w(x)dx = 0.$$

∎

Note that it is not necessary to have $u_m(a) = u_m(b) = 0$ for all m in order for the right side of Equation (33.8) to be zero; it is enough to have

$$p(a)u_m(a) = p(b)u_m(b) = 0.$$

We shall make use of this fact in our discussion of Bessel's and Legendre's equations.

33.5 Normal Form of Sturm-Liouville Equations

We can put an equation in the Sturm-Liouville form into normal form by first writing it in standard form. There is a better way, though. With the change of variable from x to μ, where

$$\mu(x) = \int_a^x \frac{1}{p(t)}dt,$$

and

$$\mu'(x) = 1/p(x),$$

we can show that

$$\frac{dy}{dx} = \frac{1}{p(x)}\frac{dy}{d\mu}$$

and

$$\frac{d^2y}{dx^2} = \frac{1}{p^2}\frac{d^2y}{d\mu^2} - \frac{p'(x)}{p(x)}\frac{dy}{d\mu}.$$

It follows that

$$\frac{d^2y}{d\mu^2} + q_1(\mu)y = 0. \tag{33.9}$$

For that reason, we study equations of the form

$$y'' + q(x)y = 0. \tag{33.10}$$

33.6 Examples

In this section we present several examples. We shall study these in more detail later in these notes.

33.6.1 Wave Equations

Separating the variables to solve wave equations leads to important ordinary differential equations.

33.6.1.1 The Homogeneous Vibrating String

The wave equation for the homogeneous vibrating string is

$$T\frac{\partial^2 u}{\partial x^2} = m\frac{\partial^2 u}{\partial t^2}, \tag{33.11}$$

where T is the constant tension and m the constant mass density. Separating the variables leads to the differential equation

$$-y''(x) = \lambda y(x). \tag{33.12}$$

33.6.1.2 The Non-homogeneous Vibrating String

When the mass density $m(x)$ varies with x, the resulting wave equation becomes

$$T\frac{\partial^2 u}{\partial x^2} = m(x)\frac{\partial^2 u}{\partial t^2}. \tag{33.13}$$

Separating the variables leads to the differential equation

$$-\frac{T}{m(x)}y''(x) = \lambda y(x). \tag{33.14}$$

33.6.1.3 The Vibrating Hanging Chain

In the hanging chain problem, considered in more detail later, the tension is not constant along the chain, since at each point it depends on the weight of the part of the chain below. The wave equation becomes

$$\frac{\partial^2 u}{\partial t^2} = g\frac{\partial}{\partial x}\Big(x\frac{\partial u}{\partial x}\Big). \tag{33.15}$$

Separating the variables leads to the differential equation

$$-g\frac{d}{dx}\Big(x\frac{dy}{dx}\Big) = \lambda y(x). \tag{33.16}$$

Note that all three of these differential equations have the form

$$Ly = \lambda y,$$

for L given by Equation (33.6).

If we make the change of variable

$$z = 2\sqrt{\frac{\lambda x}{g}},$$

the differential equation in (33.16) becomes

$$z^2\frac{d^2y}{dz^2} + z\frac{dy}{dz} + (z^2 - 0^2)y = 0. \tag{33.17}$$

As we shall see shortly, this is a special case of Bessel's Equation, with $\nu = 0$.

33.6.2 Bessel's Equations

For each non-negative constant ν the associated Bessel's Equation is

$$x^2y''(x) + xy'(x) + (x^2 - \nu^2)y(x) = 0. \tag{33.18}$$

Note that the differential equation in Equation (33.16) has the form $Ly = \lambda y$, but Equation (33.17) was obtained by a change of variable that absorbed the λ into the z, so we do not expect this form of the equation to be in eigenvalue form. However, we can rewrite Equation (33.18) as

$$-\frac{1}{x}\frac{d}{dx}\Big(xy'(x)\Big) + \frac{\nu^2}{x^2}y(x) = y(x), \tag{33.19}$$

which is in the form of a Sturm-Liouville eigenvalue problem, with $w(x) = x = p(x)$, $q(x) = \frac{\nu^2}{x^2}$, and $\lambda = 1$. As we shall discuss again in the chapter on Bessel's Equations, we can use this fact to obtain a family of orthogonal eigenfunctions.

Let us fix ν and denote by $J_\nu(x)$ a solution of Equation (33.18). Then $J_\nu(x)$ solves the eigenvalue problem in Equation (33.19), for $\lambda = 1$. A little calculation shows that for any a the function $u(x) = J_\nu(ax)$ satisfies the eigenvalue problem

$$-\frac{1}{x}\frac{d}{dx}\Big(xy'(x)\Big) + \frac{\nu^2}{x^2}y(x) = a^2y(x). \tag{33.20}$$

Let $\gamma_m > 0$ be the positive roots of $J_\nu(x)$ and define $y_m(x) = J_\nu(\gamma_m x)$ for each m. Then we have

$$-\frac{1}{x}\frac{d}{dx}\Big(xy_m'(x)\Big) + \frac{\nu^2}{x^2}y_m(x) = \gamma_m^2 y_m(x), \tag{33.21}$$

and $y_m(1) = 0$ for each m. We have the following result.

Theorem 33.3 *Let γ_m and γ_n be distinct positive zeros of $J_\nu(x)$. Then*

$$\int_0^1 y_m(x)y_n(x)xdx = 0.$$

Proof: The proof is quite similar to the proof of Theorem 33.2. The main point is that now

$$\Big(xy_n(x)y'_m(x) - xy_m(x)y'_n(x)\Big)|_0^1 = 0$$

because $y_m(1) = 0$ for all m and the function $w(x) = x$ is zero when $x = 0$.

33.6.3 Legendre's Equations

Legendre's equations have the form

$$(1 - x^2)y''(x) - 2xy'(x) + p(p + 1)y(x) = 0, \tag{33.22}$$

where p is a constant. When $p = n$ is a non-negative integer, there is a solution $P_n(x)$ that is a polynomial of degree n, containing only even or odd powers, as n is either even or odd; $P_n(x)$ is called the nth Legendre polynomial. Since the differential equation in (33.22) can be written as

$$-\frac{d}{dx}\Big((1 - x^2)y'(x)\Big) = p(p + 1)y(x), \tag{33.23}$$

it is a Sturm-Liouville eigenvalue problem with $w(x) = 1$, $p(x) = (1 - x^2)$ and $q(x) = 0$. The polynomials $P_n(x)$ are eigenfunctions of the Legendre differential operator T given by

$$(Ty)(x) = -\frac{d}{dx}\Big((1 - x^2)y'(x)\Big), \tag{33.24}$$

but we have not imposed any explicit boundary conditions. Nevertheless, we have the following orthogonality theorem.

Theorem 33.4 *For $m \neq n$ we have*

$$\int_{-1}^1 P_m(x)P_n(x)dx = 0.$$

Proof: In this case, Equation (33.8) becomes

$$(\lambda_n - \lambda_m)\int_{-1}^1 P_m(x)P_n(x)dx =$$

$$\Big((1 - x^2)[P_n(x)P'_m(x) - P_m(x)P'_n(x)]\Big)|_{-1}^1 = 0, \tag{33.25}$$

which holds not because we have imposed end-point conditions on the $P_n(x)$, but because $p(x) = 1 - x^2$ is zero at both ends. ∎

33.6.4 Other Famous Examples

Well known examples of Sturm-Liouville problems also include

- **Chebyshev:**

$$\frac{d}{dx}\left(\sqrt{1-x^2}\frac{dy}{dx}\right) + \lambda(1-x^2)^{-1/2}y = 0;$$

- **Hermite:**

$$\frac{d}{dx}\left(e^{-x^2}\frac{dy}{dx}\right) + \lambda e^{-x^2}y = 0;$$

and

- **Laguerre:**

$$\frac{d}{dx}\left(xe^{-x}\frac{dy}{dx}\right) + \lambda e^{-x}y = 0.$$

Ex. 33.1 *For each of the three differential equations just listed, see if you can determine the interval over which their eigenfunctions will be orthogonal.*

In the next appendix we consider Hermite's Equation and its connection to quantum mechanics.

Chapter 34

Appendix: Hermite's Equations and Quantum Mechanics

34.1 The Schrödinger Wave Function

In quantum mechanics, the behavior of a particle with mass m subject to a potential $V(x,t)$ satisfies the Schrödinger Equation

$$i\hbar\frac{\partial\psi(x,t)}{\partial t} = -\frac{\hbar}{2m}\frac{\partial^2\psi(x,t)}{\partial x^2} + V(x,t)\psi(x,t), \tag{34.1}$$

where $\hbar$ is Planck's constant. Here the x is one-dimensional, but extensions to higher dimensions are also possible.

When the solution $\psi(x,t)$ is selected so that

$$|\psi(x,t)| \to 0,$$

as $|x| \to \infty$, and

$$\int_{-\infty}^{\infty} |\psi(x,t)|^2 dx = 1,$$

then, for each fixed t, the function $|\psi(x,t)|^2$ is a probability density function governing the position of the particle. In other words, the probability of finding the particle in the interval $[a,b]$ at time t is

$$\int_a^b |\psi(x,t)|^2 dx.$$

An important special case is that of time-independent potentials.

34.2 Time-Independent Potentials

We say that $V(x,t)$ is time-independent if $V(x,t) = V(x)$, for all t. We then attempt to solve Equation (34.1) by separating the variables; we take $\psi(x,t) = f(t)g(x)$ and insert this product into Equation (34.1).

The time function is easily shown to be

$$f(t) = e^{-Et/\hbar},$$

where E is defined to be the energy. The function $g(x)$ satisfies the *time-independent Schrödinger Equation*

$$-\frac{\hbar}{2m}g''(x) + V(x)g(x) = Eg(x). \tag{34.2}$$

An important special case is the harmonic oscillator.

34.3 The Harmonic Oscillator

The case of the *harmonic oscillator* corresponds to the potential $V(x) = \frac{1}{2}kx^2$.

34.3.1 The Classical Spring Problem

To motivate the development of the harmonic oscillator in quantum mechanics, it is helpful to recall the classical spring problem. In this problem a mass m slides back and forth along a frictionless surface, with position $x(t)$ at time t. It is connected to a fixed structure by a spring with spring constant $k > 0$. The restoring force acting on the mass at any time is $-kx$, with $x = 0$ the equilibrium position of the mass. The equation of motion is

$$mx''(t) = -kx(t),$$

and the solution is

$$x(t) = x(0)\cos\sqrt{\frac{k}{m}}t.$$

The period of oscillation is $T = 2\pi\sqrt{\frac{m}{k}}$ and the frequency of oscillation is $\nu = \frac{1}{T} = \frac{1}{2\pi}\sqrt{\frac{k}{m}}$, from which we obtain the equation

$$k = 4\pi^2 m\nu^2.$$

The potential energy is $\frac{1}{2}kx^2$, while the kinetic energy is $\frac{1}{2}m\dot{x}^2$. The sum of the kinetic and potential energies is the total energy, $E(t)$. Since $E'(t) = 0$, the energy is constant.

34.3.2 Back to the Harmonic Oscillator

When the potential function is $V(x) = \frac{1}{2}kx^2$, Equation (34.2) becomes

$$\frac{\hbar}{2m}g''(x) + (E - \frac{1}{2}kx^2)g(x) = 0, \tag{34.3}$$

where $k = m\omega^2$, for $\omega = 2\pi\nu$. With $u = \sqrt{\frac{m\omega}{\hbar}}$ and $\epsilon = \frac{2E}{\hbar\omega}$, we have

$$\frac{d^2g}{du^2} + (\epsilon - u^2)g = 0. \tag{34.4}$$

Equation (34.4) is equivalent to

$$w''(x) + (2p + 1 - x^2)w(x) = 0,$$

which can be transformed into Hermite's Equation

$$y'' - 2xy' + 2py = 0,$$

by writing $y(x) = w(x)e^{x^2/2}$.

In order for the solutions of Equation (34.3) to be physically admissible solutions, it is necessary that p be a non-negative integer, which means that

$$E = \hbar\omega(n + \frac{1}{2}),$$

for some non-negative integer n; this gives the *quantized energy levels* for the harmonic oscillator.

34.4 Dirac's Equation

Einstein's theory of special relativity tells us that there are four variables, not just three, that have length for their units of measurement: the familiar three-dimensional spatial coordinates, and ct, where c is the speed of light and t is time. Looked at this way, Schrödinger's Equation (34.1), extended to three spatial dimensions, is peculiar, in that it treats the variable ct differently from the others. There is only a first partial derivative in t, but second partial derivatives in the other variables. In 1930 the British

mathematician Paul Dirac presented his relativistically correct version of Schrödinger's Equation.

Dirac's Equation, a version of which is inscribed on the wall of Westminster Abbey, is the following:

$$i\hbar\frac{\partial\psi}{\partial t} = \frac{\hbar c}{i}\Big(\alpha_1\frac{\partial\psi}{\partial x_1} + \alpha_2\frac{\partial\psi}{\partial x_2} + \alpha_3\frac{\partial\psi}{\partial x_3}\Big) + \alpha_4 mc^2\psi. \tag{34.5}$$

Here the α_i are the Dirac matrices.

This equation agreed remarkably well with experimental data on the behavior of electrons in electric and magnetic fields, but it also seemed to allow for nonsensical solutions, such as spinning electrons with negative energy. The next year, Dirac realized that what the equation was calling for was *anti-matter*, a particle with the same mass as the electron, but with a positive charge. In the summer of 1932 Carl Anderson, working at Cal Tech, presented clear evidence for the existence of such a particle, which we now call the *positron*. What seemed like the height of science fiction in 1930 has become commonplace today.

When a positron collides with an electron their masses vanish and two gamma ray photons of pure energy are produced. These photons then move off in opposite directions. In positron emission tomography (PET) certain positron-emitting chemicals, such as glucose with radioactive fluorine chemically attached, are injected into the patient. When the PET scanner detects two photons arriving at the two ends of a line segment at (almost) the same time, called *coincidence detection*, it concludes that a positron was emitted somewhere along that line. This is repeated thousands of times. Once all this data has been collected, the mathematicians take over and use these clues to reconstruct an image of where the glucose is in the body. It is this image that the doctor sees.

Chapter 35

Appendix: The BLUE and The Kalman Filter

35.1 Chapter Summary

In most signal- and image-processing applications the measured data includes (or may include) a signal component we want and unwanted components called *noise*. Estimation involves determining the precise nature and strength of the signal component; deciding if that strength is zero or not is detection.

Noise often appears as an additive term, which we then try to remove. If we knew precisely the noisy part added to each data value we would simply subtract it; of course, we never have such information. How then do we remove something when we don't know what it is? Statistics provides a way out.

The basic idea in statistics is to use procedures that perform well on average, when applied to a class of problems. The procedures are built using properties of that class, usually involving probabilistic notions, and are evaluated by examining how they would have performed had they been applied to every problem in the class. To use such methods to remove additive noise, we need a description of the class of noises we expect to encounter, not specific values of the noise component in any one particular instance. We also need some idea about what signal components look like. In this chapter we discuss solving this noise removal problem using the *best*

linear unbiased estimation (BLUE). We begin with the simplest case and then proceed to discuss increasingly complex scenarios.

An important application of the BLUE is in Kalman filtering. The connection between the BLUE and Kalman filtering is best understood by considering the case of the BLUE with a prior estimate of the signal component, and mastering the various matrix manipulations that are involved in this problem. These calculations then carry over, almost unchanged, to the Kalman filtering.

Kalman filtering is usually presented in the context of estimating a sequence of vectors evolving in time. Kalman filtering for image processing is derived by analogy with the temporal case, with certain parts of the image considered to be in the "past" of a fixed pixel.

35.2 The Simplest Case

Suppose our data is $z_j = c + v_j$, for $j = 1, ..., J$, where c is an unknown constant to be estimated and the v_j are additive noise. We assume that $E(v_j) = 0, E(v_j\overline{v_k}) = 0$ for $j \neq k$, and $E(|v_j|^2) = \sigma_j^2$. So, the additive noises are assumed to have mean zero and to be independent (or at least uncorrelated). In order to estimate c, we adopt the following rules:

- **1.** The estimate $\hat{c}$ is *linear* in the data $\mathbf{z} = (z_1, ..., z_J)^T$; that is, $\hat{c} = \mathbf{k}^\dagger \mathbf{z}$, for some vector $\mathbf{k} = (k_1, ..., k_J)^T$.

- **2.** The estimate is *unbiased*; $E(\hat{c}) = c$. This means $\sum_{j=1}^J k_j = 1$.

- **3.** The estimate is best in the sense that it minimizes the expected error squared; that is, $E(|\hat{c} - c|^2)$ is minimized.

Ex. 35.1 *Show that the resulting vector* $\mathbf{k}$ *is*

$$k_i = \sigma_i^{-2} / (\sum_{j=1}^J \sigma_j^{-2}),$$

and the BLUE estimator of c *is then*

$$\hat{c} = \sum_{i=1}^J z_i \sigma_i^{-2} / (\sum_{j=1}^J \sigma_j^{-2}).$$

Ex. 35.2 *Suppose we have data* $z_1 = c + v_1$ *and* $z_2 = c + v_2$ *and we want to estimate the constant* c*. Assume that* $E(v_1) = E(v_2) = 0$ *and* $E(v_1 v_2) = \rho$*, with* $0 < |\rho| < 1$*. Find the BLUE estimate of* c*.*

Ex. 35.3 *The concentration of a substance in solution decreases exponentially during an experiment. Noisy measurements of the concentration are made at times t_1 and t_2, giving the data*

$$z_i = x_0 e^{-t_i} + v_i, \ i = 1, 2,$$

where the v_i have mean zero, and are uncorrelated. Find the BLUE for the initial concentration x_0.

35.3 A More General Case

Suppose now that our data vector is $\mathbf{z} = H\mathbf{x} + \mathbf{v}$. Here, $\mathbf{x}$ is an unknown vector whose value is to be estimated, the random vector $\mathbf{v}$ is additive noise whose mean is $E(\mathbf{v}) = 0$ and whose known covariance matrix is $Q = E(\mathbf{v}\mathbf{v}^\dagger)$, not necessarily diagonal, and the known matrix H is J by N, with $J > N$. Now we seek an estimate of the vector $\mathbf{x}$. We now use the following rules:

- **1.** The estimate $\hat{\mathbf{x}}$ must have the form $\hat{\mathbf{x}} = K^\dagger \mathbf{z}$, where the matrix K is to be determined.
- **2.** The estimate is unbiased; that is, $E(\hat{\mathbf{x}}) = \mathbf{x}$.
- **3.** The K is determined as the minimizer of the expected squared error; that is, once again we minimize $E(|\hat{\mathbf{x}} - \mathbf{x}|^2)$.

Ex. 35.4 *Show that for the estimator to be unbiased we need $K^\dagger H = I$, the identity matrix.*

Ex. 35.5 *Show that*

$$E(|\hat{\mathbf{x}} - \mathbf{x}|^2) = \text{trace}\, K^\dagger Q K.$$

Hints: Write the left side as

$$E(\text{trace}\,((\hat{\mathbf{x}} - \mathbf{x})(\hat{\mathbf{x}} - \mathbf{x})^\dagger)).$$

Also use the fact that the trace and expected-value operations commute.

The problem then is to minimize trace $K^\dagger QK$ subject to the constraint equation $K^\dagger H = I$. We solve this problem using a technique known as *prewhitening*.

Since the noise covariance matrix Q is Hermitian and nonnegative definite, we have $Q = UDU^\dagger$, where the columns of U are the (mutually orthogonal) eigenvectors of Q and D is a diagonal matrix whose diagonal entries are the (necessarily nonnegative) eigenvalues of Q; therefore, $U^\dagger U = I$. We call $C = UD^{1/2}U^\dagger$ the Hermitian square root of Q, since $C^\dagger = C$ and $C^2 = Q$. We assume that Q is invertible, so that C is also. Given the system of equations

$$\mathbf{z} = H\mathbf{x} + \mathbf{v},$$

as before, we obtain a new system

$$\mathbf{y} = G\mathbf{x} + \mathbf{w}$$

by multiplying both sides by $C^{-1} = Q^{-1/2}$; here, $G = C^{-1}H$ and $\mathbf{w} = C^{-1}\mathbf{v}$. The new noise correlation matrix is

$$E(\mathbf{w}\mathbf{w}^\dagger) = C^{-1}QC^{-1} = I,$$

so the new noise is white. For this reason the step of multiplying by C^{-1} is called *prewhitening*.

With $J = CK$ and $M = C^{-1}H$, we have

$$K^\dagger QK = J^\dagger J$$

and

$$K^\dagger H = J^\dagger M.$$

Our problem then is to minimize trace $J^\dagger J$, subject to $J^\dagger M = I$. Recall that the trace of the matrix $A^\dagger A$ is simply the square of the 2-norm of the vectorization of A.

Our solution method is to transform the original problem into a simpler problem, where the answer is obvious.

First, for any given matrices L and M such that J and ML have the same dimensions, the minimum value of

$$f(J) = \text{trace}[(J^\dagger - L^\dagger M^\dagger)(J - ML)]$$

is zero and occurs when $J = ML$.

Now let $L = L^\dagger = (M^\dagger M)^{-1}$. The solution is again $J = ML$, but now this choice for J has the additional property that $J^\dagger M = I$. So, minimizing $f(J)$ is equivalent to minimizing $f(J)$ subject to the constraint $J^\dagger M = I$ and both problems have the solution $J = ML$.

Now using $J^\dagger M = I$, we expand $f(J)$ to get

$$f(J) = \text{trace}[J^\dagger J - J^\dagger ML - L^\dagger M^\dagger J + L^\dagger M^\dagger ML]$$

$$= \text{trace}[J^\dagger J - L - L^\dagger + L^\dagger M^\dagger ML].$$

The only term here that involves the unknown matrix J is the first one. Therefore, minimizing $f(J)$ subject to $J^\dagger M = I$ is equivalent to minimizing trace $J^\dagger J$ subject to $J^\dagger M = I$, which is our original problem. Therefore, the optimal choice for J is $J = ML$. Consequently, the optimal choice for K is

$$K = Q^{-1}HL = Q^{-1}H(H^\dagger Q^{-1}H)^{-1},$$

and the BLUE estimate of $\mathbf{x}$ is

$$\mathbf{x}_{BLUE} = \hat{\mathbf{x}} = K^\dagger \mathbf{z} = (H^\dagger Q^{-1}H)^{-1}H^\dagger Q^{-1}\mathbf{z}.$$

The simplest case can be obtained from this more general formula by taking $N = 1$, $H = (1, 1, ..., 1)^T$ and $\mathbf{x} = c$.

Note that if the noise is *white*, that is, $Q = \sigma^2 I$, then $\hat{\mathbf{x}} = (H^\dagger H)^{-1}H^\dagger \mathbf{z}$, which is the least-squares solution of the equation $\mathbf{z} = H\mathbf{x}$. The effect of requiring that the estimate be unbiased is that, in this case, we simply ignore the presence of the noise and calculate the least squares solution of the noise-free equation $\mathbf{z} = H\mathbf{x}$.

The BLUE estimator involves nested inversion, making it difficult to calculate, especially for large matrices. In the exercise that follows, we discover an approximation of the BLUE that is easier to calculate.

Ex. 35.6 *Show that for $\epsilon > 0$ we have*

$$(H^\dagger Q^{-1}H + \epsilon I)^{-1}H^\dagger Q^{-1} = H^\dagger (HH^\dagger + \epsilon Q)^{-1}. \tag{35.1}$$

Hint: Use the identity

$$H^\dagger Q^{-1}(HH^\dagger + \epsilon Q) = (H^\dagger Q^{-1}H + \epsilon I)H^\dagger.$$

It follows from Equation (35.1) that

$$\mathbf{x}_{BLUE} = \lim_{\epsilon \to 0} H^\dagger (HH^\dagger + \epsilon Q)^{-1}\mathbf{z}. \tag{35.2}$$

Therefore, we can get an approximation of the BLUE estimate by selecting $\epsilon > 0$ near zero, solving the system of linear equations

$$(HH^\dagger + \epsilon Q)\mathbf{a} = \mathbf{z}$$

for $\mathbf{a}$ and taking $\mathbf{x} = H^\dagger \mathbf{a}$.

35.4 Some Useful Matrix Identities

In the exercise that follows we consider several matrix identities that are useful in developing the Kalman filter.

Ex. 35.7 *Establish the following identities, assuming that all the products and inverses involved are defined:*

$$CDA^{-1}B(C^{-1} - DA^{-1}B)^{-1} = (C^{-1} - DA^{-1}B)^{-1} - C; \tag{35.3}$$

$$(A - BCD)^{-1} = A^{-1} + A^{-1}B(C^{-1} - DA^{-1}B)^{-1}DA^{-1}; \tag{35.4}$$

$$A^{-1}B(C^{-1} - DA^{-1}B)^{-1} = (A - BCD)^{-1}BC; \tag{35.5}$$

$$(A - BCD)^{-1} = (I + GD)A^{-1}, \tag{35.6}$$

for

$$G = A^{-1}B(C^{-1} - DA^{-1}B)^{-1}.$$

Hints: To get Equation (35.3) use

$$C(C^{-1} - DA^{-1}B) = I - CDA^{-1}B.$$

For the second identity, multiply both sides of Equation (35.4) on the left by $A - BCD$ and at the appropriate step use Equation (35.3). For Equation (35.5) show that

$$BC(C^{-1} - DA^{-1}B) = B - BCDA^{-1}B = (A - BCD)A^{-1}B.$$

For Equation (35.6), substitute what G is and use Equation (35.4).

35.5 The BLUE with a Prior Estimate

In Kalman filtering we have the situation in which we want to estimate an unknown vector $\mathbf{x}$ given measurements $\mathbf{z} = H\mathbf{x} + \mathbf{v}$, but also given a prior estimate $\mathbf{y}$ of $\mathbf{x}$. It is the case there that $E(\mathbf{y}) = E(\mathbf{x})$, so we write $\mathbf{y} = \mathbf{x} + \mathbf{w}$, with $\mathbf{w}$ independent of both $\mathbf{x}$ and $\mathbf{v}$ and $E(\mathbf{w}) = \mathbf{0}$. The

covariance matrix for $\mathbf{w}$ we denote by $E(\mathbf{w}\mathbf{w}^\dagger) = R$. We now require that the estimate $\hat{\mathbf{x}}$ be linear in both $\mathbf{z}$ and $\mathbf{y}$; that is, the estimate has the form

$$\hat{\mathbf{x}} = C^\dagger \mathbf{z} + D^\dagger \mathbf{y},$$

for matrices C and D to be determined.

The approach is to apply the BLUE to the combined system of linear equations

$$\mathbf{z} = H\mathbf{x} + \mathbf{v} \quad \text{and}$$

$$\mathbf{y} = \mathbf{x} + \mathbf{w}.$$

In matrix language this combined system becomes $\mathbf{u} = J\mathbf{x} + \mathbf{n}$, with $\mathbf{u}^T = [\mathbf{z}^T\ \mathbf{y}^T]$, $J^T = [H^T\ I^T]$, and $\mathbf{n}^T = [\mathbf{v}^T\ \mathbf{w}^T]$. The noise covariance matrix becomes

$$P = \begin{bmatrix} Q & 0 \\ 0 & R \end{bmatrix}.$$

The BLUE estimate is $K^\dagger \mathbf{u}$, with $K^\dagger J = I$. Minimizing the variance, we find that the optimal $K^\dagger$ is

$$K^\dagger = (J^\dagger P^{-1} J)^{-1} J^\dagger P^{-1}.$$

The optimal estimate is then

$$\hat{\mathbf{x}} = (H^\dagger Q^{-1} H + R^{-1})^{-1} (H^\dagger Q^{-1} \mathbf{z} + R^{-1} \mathbf{y}).$$

Therefore,

$$C^\dagger = (H^\dagger Q^{-1} H + R^{-1})^{-1} H^\dagger Q^{-1}$$

and

$$D^\dagger = (H^\dagger Q^{-1} H + R^{-1})^{-1} R^{-1}.$$

Using the matrix identities in Equations (35.4) and (35.5) we can rewrite this estimate in the more useful form

$$\hat{\mathbf{x}} = \mathbf{y} + G(\mathbf{z} - H\mathbf{y}),$$

for

$$G = RH^\dagger (Q + HRH^\dagger)^{-1}. \tag{35.7}$$

The covariance matrix of the optimal estimator is $K^\dagger P K$, which can be written as

$$K^\dagger P K = (R^{-1} + H^\dagger Q^{-1} H)^{-1} = (I - GH)R.$$

In the context of the Kalman filter, R is the covariance of the prior estimate of the current state, G is the Kalman gain matrix, and $K^\dagger P K$ is the posterior covariance of the current state. The algorithm proceeds recursively from one state to the next in time.

35.6 Adaptive BLUE

We have assumed so far that we know the covariance matrix Q corresponding to the measurement noise. If we do not, then we may attempt to estimate Q from the measurements themselves; such methods are called *noise-adaptive*. To illustrate, let the *innovations* vector be $\mathbf{e} = \mathbf{z} - H\mathbf{y}$. Then the covariance matrix of $\mathbf{e}$ is $S = HRH^\dagger + Q$. Having obtained an estimate $\hat{S}$ of S from the data, we use $\hat{S} - HRH^\dagger$ in place of Q in Equation (35.7).

35.7 The Kalman Filter

So far in this chapter we have focused on the filtering problem: given the data vector $\mathbf{z}$, estimate $\mathbf{x}$, assuming that $\mathbf{z}$ consists of noisy measurements of $H\mathbf{x}$; that is, $\mathbf{z} = H\mathbf{x} + \mathbf{v}$. An important extension of this problem is that of stochastic prediction. Shortly, we discuss the Kalman-filter method for solving this more general problem. One area in which prediction plays an important role is the tracking of moving targets, such as ballistic missiles, using radar. The range to the target, its angle of elevation, and its azimuthal angle are all functions of time governed by linear differential equations. The *state vector* of the system at time t might then be a vector with nine components, the three functions just mentioned, along with their first and second derivatives. In theory, if we knew the initial state perfectly and our differential equations model of the physics was perfect, that would be enough to determine the future states. In practice neither of these is true, and we need to assist the differential equation by taking radar measurements of the state at various times. The problem then is to estimate the state at time t using both the measurements taken prior to time t and the estimate based on the physics.

When such tracking is performed digitally, the functions of time are replaced by discrete sequences. Let the state vector at time $k\Delta t$ be denoted by $\mathbf{x}_k$, for k an integer and $\Delta t > 0$. Then, with the derivatives in the differential equation approximated by divided differences, the physical model for the evolution of the system in time becomes

$$\mathbf{x}_k = A_{k-1}\mathbf{x}_{k-1} + \mathbf{m}_{k-1}.$$

The matrix A_{k-1}, which we assume is known, is obtained from the differential equation, which may have nonconstant coefficients, as well as from the

divided difference approximations to the derivatives. The random vector sequence $\mathbf{m}_{k-1}$ represents the error in the physical model due to the discretization and necessary simplification inherent in the original differential equation itself. We assume that the expected value of $\mathbf{m_k}$ is zero for each k. The covariance matrix is $E(\mathbf{m}_k\mathbf{m}_k^\dagger) = M_k$.

At time $k\Delta t$ we have the measurements

$$\mathbf{z}_k = H_k\mathbf{x}_k + \mathbf{v}_k,$$

where H_k is a known matrix describing the nature of the linear measurements of the state vector and the random vector $\mathbf{v}_k$ is the noise in these measurements. We assume that the mean value of $\mathbf{v}_k$ is zero for each k. The covariance matrix is $E(\mathbf{v}_k\mathbf{v}_k^\dagger) = Q_k$. We assume that the initial state vector $\mathbf{x}_0$ is arbitrary.

Given an unbiased estimate $\hat{\mathbf{x}}_{k-1}$ of the state vector $\mathbf{x}_{k-1}$, our prior estimate of $\mathbf{x}_k$ based solely on the physics is

$$\mathbf{y}_k = A_{k-1}\hat{\mathbf{x}}_{k-1}.$$

Ex. 35.8 *Show that $E(\mathbf{y}_k - \mathbf{x}_k) = 0$, so the prior estimate of $\mathbf{x}_k$ is unbiased. We can then write $\mathbf{y}_k = \mathbf{x}_k + \mathbf{w}_k$, with $E(\mathbf{w}_k) = \mathbf{0}$.*

35.8 Kalman Filtering and the BLUE

The *Kalman filter* [178, 143, 93] is a recursive algorithm to estimate the state vector $\mathbf{x}_k$ at time $k\Delta t$ as a linear combination of the vectors $\mathbf{z}_k$ and $\mathbf{y}_k$. The estimate $\hat{\mathbf{x}}_k$ will have the form

$$\hat{\mathbf{x}}_k = C_k^\dagger\mathbf{z}_k + D_k^\dagger\mathbf{y}_k, \tag{35.8}$$

for matrices C_k and D_k to be determined. As we shall see, this estimate can also be written as

$$\hat{\mathbf{x}}_k = \mathbf{y}_k + G_k(\mathbf{z}_k - H_k\mathbf{y}_k), \tag{35.9}$$

which shows that the estimate involves a prior prediction step, the $\mathbf{y}_k$, followed by a correction step, in which $H_k\mathbf{y}_k$ is compared to the measured data vector $\mathbf{z}_k$; such estimation methods are sometimes called *predictor-corrector methods*.

In our discussion of the BLUE, we saw how to incorporate a prior estimate of the vector to be estimated. The trick was to form a larger

matrix equation and then to apply the BLUE to that system. The Kalman filter does just that.

The correction step in the Kalman filter uses the BLUE to solve the combined linear system

$$\mathbf{z}_k = H_k\mathbf{x}_k + \mathbf{v}_k$$

and

$$\mathbf{y}_k = \mathbf{x}_k + \mathbf{w}_k.$$

The covariance matrix of $\hat{\mathbf{x}}_{k-1} - \mathbf{x}_{k-1}$ is denoted by P_{k-1}, and we let $Q_k = E(\mathbf{w}_k\mathbf{w}_k^\dagger)$. The covariance matrix of $\mathbf{y}_k - \mathbf{x}_k$ is

$$\text{cov}(\mathbf{y}_k - \mathbf{x}_k) = R_k = M_{k-1} + A_{k-1}P_{k-1}A_{k-1}^\dagger.$$

It follows from our earlier discussion of the BLUE that the estimate of $\mathbf{x}_k$ is

$$\hat{\mathbf{x}}_k = \mathbf{y}_k + G_k(\mathbf{z}_k - H\mathbf{y}_k),$$

with

$$G_k = R_kH_k^\dagger(Q_k + H_kR_kH_k^\dagger)^{-1}.$$

Then, the covariance matrix of $\hat{\mathbf{x}}_k - \mathbf{x}_k$ is

$$P_k = (I - G_kH_k)R_k.$$

The recursive procedure is to go from P_{k-1} and M_{k-1} to R_k, then to G_k, from which $\hat{\mathbf{x}}_k$ is formed, and finally to P_k, which, along with the known matrix M_k, provides the input to the next step. The time-consuming part of this recursive algorithm is the matrix inversion in the calculation of G_k. Simpler versions of the algorithm are based on the assumption that the matrices Q_k are diagonal, or on the convergence of the matrices G_k to a limiting matrix G [93].

There are many variants of the Kalman filter, corresponding to variations in the physical model, as well as in the statistical assumptions. The differential equation may be nonlinear, so that the matrices A_k depend on $\mathbf{x}_k$. The system noise sequence $\{\mathbf{w}_k\}$ and the measurement noise sequence $\{\mathbf{v}_k\}$ may be correlated. For computational convenience the various functions that describe the state may be treated separately. The model may include known external inputs to drive the differential system, as in the tracking of spacecraft capable of firing booster rockets. Finally, the noise covariance matrices may not be known *a priori* and adaptive filtering may be needed. We discuss this last issue briefly in the next section.

35.9 Adaptive Kalman Filtering

As in [93] we consider only the case in which the covariance matrix Q_k of the measurement noise $\mathbf{v}_k$ is unknown. As we saw in the discussion of adaptive BLUE, the covariance matrix of the innovations vector $\mathbf{e}_k = \mathbf{z}_k - H_k\mathbf{y}_k$ is

$$S_k = H_k R_k H_k^\dagger + Q_k.$$

Once we have an estimate for S_k, we estimate Q_k using

$$\hat{Q}_k = \hat{S}_k - H_k R_k H_k^\dagger.$$

We might assume that S_k is independent of k and estimate $S_k = S$ using past and present innovations; for example, we could use

$$\hat{S} = \frac{1}{k-1}\sum_{j=1}^{k}(\mathbf{z}_j - H_j\mathbf{y}_j)(\mathbf{z}_j - H_j\mathbf{y}_j)^\dagger.$$

Chapter 36

Appendix: Matrix and Vector Differentiation

36.1 Chapter Summary

The notation associated with matrix and vector algebra is designed to reduce the number of things we have to think about as we perform our calculations. This notation can be extended to multi-variable calculus, as we show in this chapter.

36.2 Functions of Vectors and Matrices

As we saw in the previous chapter, the least squares approximate solution of $A\mathbf{x} = \mathbf{b}$ is a vector $\hat{\mathbf{x}}$ that minimizes the function $||A\mathbf{x} - \mathbf{b}||$. In our discussion of band-limited extrapolation we showed that, for any nonnegative definite matrix Q, the vector having norm one that maximizes the quadratic form $\mathbf{x}^\dagger Q\mathbf{x}$ is an eigenvector of Q associated with the largest eigenvalue. In the chapter on best linear unbiased optimization we seek a matrix that minimizes a certain function. All of these examples involve what we can call *matrix-vector differentiation*, that is, the differentiation of a function with respect to a matrix or a vector. The gradient of a function of several variables is a well-known example and we begin there. Since there is some possibility of confusion, we adopt the notational convention that boldfaced symbols, such as $\mathbf{x}$, indicate a column vector, while x denotes a scalar.

36.3 Differentiation with Respect to a Vector

Let $\mathbf{x} = (x_1, ..., x_N)^T$ be an N-dimensional real column vector. Let $z = f(\mathbf{x})$ be a real-valued function of the entries of $\mathbf{x}$. The derivative of z with respect to $\mathbf{x}$, also called the *gradient* of z, is the column vector

$$\frac{\partial z}{\partial \mathbf{x}} = \mathbf{a} = (a_1, ..., a_N)^T$$

with entries

$$a_n = \frac{\partial z}{\partial x_n}.$$

Ex. 36.1 *Let* $\mathbf{y}$ *be a fixed real column vector and* $z = f(\mathbf{x}) = \mathbf{y}^T\mathbf{x}$. *Show that*

$$\frac{\partial z}{\partial \mathbf{x}} = \mathbf{y}.$$

Ex. 36.2 *Let* Q *be a real symmetric nonnegative definite matrix, and let* $z = f(\mathbf{x}) = \mathbf{x}^T Q\mathbf{x}$. *Show that the gradient of this quadratic form is*

$$\frac{\partial z}{\partial \mathbf{x}} = 2Q\mathbf{x}.$$

Hint: Write Q as a linear combination of dyads involving the eigenvectors.

Ex. 36.3 *Let* $z = ||A\mathbf{x} - \mathbf{b}||^2$. *Show that*

$$\frac{\partial z}{\partial \mathbf{x}} = 2A^T A\mathbf{x} - 2A^T\mathbf{b}.$$

Hint: Use $z = (A\mathbf{x} - \mathbf{b})^T(A\mathbf{x} - \mathbf{b})$.

We can also consider the second derivative of $z = f(\mathbf{x})$, which is the *Hessian matrix* of z

$$H = \frac{\partial^2 z}{\partial \mathbf{x}^2} = \nabla^2 f(x)$$

with entries

$$H_{mn} = \frac{\partial^2 z}{\partial x_m \partial x_n}.$$

If the entries of the vector $\mathbf{z} = (z_1, ..., z_M)^T$ are real-valued functions of the vector $\mathbf{x}$, the derivative of $\mathbf{z}$ is the matrix whose mth column is the derivative of the real-valued function z_m. This matrix is usually called the *Jacobian matrix* of $\mathbf{z}$. If $M = N$ the determinant of the Jacobian matrix is the *Jacobian*.

Ex. 36.4 *Suppose* $(u, v) = (u(x, y), v(x, y))$ *is a change of variables from the Cartesian* (x, y) *coordinate system to some other* (u, v) *coordinate system. Let* $\mathbf{x} = (x, y)^T$ *and* $\mathbf{z} = (u(\mathbf{x}), v(\mathbf{x}))^T$.

- **(a)** *Calculate the Jacobian for the rectangular coordinate system obtained by rotating the* (x, y) *system through an angle of* θ.
- **(b)** *Calculate the Jacobian for the transformation from the* (x, y) *system to polar coordinates.*

36.4 Differentiation with Respect to a Matrix

Now we consider real-valued functions $z = f(A)$ of a real matrix A. As an example, for square matrices A we have

$$z = f(A) = \text{trace}\,(A) = \sum_{n=1}^{N} A_{nn},$$

the sum of the entries along the main diagonal of A.

The derivative of $z = f(A)$ is the matrix

$$\frac{\partial z}{\partial A} = B$$

whose entries are

$$B_{mn} = \frac{\partial z}{\partial A_{mn}}.$$

Ex. 36.5 *Show that the derivative of* $\text{trace}\,(A)$ *is* $B = I$*, the identity matrix.*

Ex. 36.6 *Show that the derivative of* $z = \text{trace}\,(DAC)$ *with respect to* A *is*

$$\frac{\partial z}{\partial A} = D^T C^T. \tag{36.1}$$

Consider the function f defined for all J by J positive-definite symmetric matrices by

$$f(Q) = -\log\det(Q). \tag{36.2}$$

Proposition 36.1 *The gradient of $f(Q)$ is $g(Q) = -Q^{-1}$.*

Proof: Let ΔQ be symmetric. Let γ_j, for $j = 1, 2, ..., J$, be the eigenvalues of the symmetric matrix $Q^{-1/2}(\Delta Q)Q^{-1/2}$. These γ_j are then real and are also the eigenvalues of the matrix $Q^{-1}(\Delta Q)$. We shall consider $\|\Delta Q\|$ small, so we may safely assume that $1 + \gamma_j > 0$.

Note that

$$\langle Q^{-1}, \Delta Q\rangle = \sum_{j=1}^{J} \gamma_j,$$

since the trace of any square matrix is the sum of its eigenvalues. Then we have

$$f(Q + \Delta Q) - f(Q) = -\log\det(Q + \Delta Q) + \log\det(Q)$$

$$= -\log\det(I + Q^{-1}(\Delta Q)) = -\sum_{j=1}^{J} \log(1 + \gamma_j).$$

From the submultiplicativity of the Frobenius norm we have

$$\|Q^{-1}(\Delta Q)\|/\|Q^{-1}\| \leq \|\Delta Q\| \leq \|Q^{-1}(\Delta Q)\|\|Q\|.$$

Therefore, taking the limit as $\|\Delta Q\|$ goes to zero is equivalent to taking the limit as $\|\gamma\|$ goes to zero, where γ is the vector whose entries are the γ_j.

To show that $g(Q) = -Q^{-1}$ note that

$$\limsup_{\|\Delta Q\|\to 0} \frac{f(Q + \Delta Q) - f(Q) - \langle -Q^{-1}, \Delta Q\rangle}{\|\Delta Q\|}$$

$$= \limsup_{\|\Delta Q\|\to 0} \frac{|-\log\det(Q + \Delta Q) + \log\det(Q) + \langle Q^{-1}, \Delta Q\rangle|}{\|\Delta Q\|}$$

$$\leq \limsup_{\|\gamma\|\to 0} \frac{\sum_{j=1}^{J} |\log(1 + \gamma_j) - \gamma_j|}{\|\gamma\|/\|Q^{-1}\|}$$

$$\leq \|Q^{-1}\| \sum_{j=1}^{J} \lim_{\gamma_j\to 0} \frac{\gamma_j - \log(1 + \gamma_j)}{|\gamma_j|} = 0.$$

∎

We note in passing that the derivative of $\det(DAC)$ with respect to A is the matrix $\det(DAC)(A^{-1})^T$.

Although the trace is not independent of the order of the matrices in a product, it is independent of cyclic permutation of the factors:

$$\text{trace}\,(ABC) = \text{trace}\,(CAB) = \text{trace}\,(BCA).$$

Therefore, the trace is independent of the order for the product of two matrices:

$$\text{trace}\,(AB) = \text{trace}\,(BA).$$

From this fact we conclude that

$$\mathbf{x}^T\mathbf{x} = \text{trace}\,(\mathbf{x}^T\mathbf{x}) = \text{trace}\,(\mathbf{x}\mathbf{x}^T).$$

If $\mathbf{x}$ is a random vector with correlation matrix

$$R = E(\mathbf{x}\mathbf{x}^T),$$

then

$$E(\mathbf{x}^T\mathbf{x}) = E(\text{trace}\,(\mathbf{x}\mathbf{x}^T)) = \text{trace}\,(E(\mathbf{x}\mathbf{x}^T)) = \text{trace}\,(R).$$

We shall use this trick in the chapter on detection.

Ex. 36.7 *Let* $z = \text{trace}\,(A^TCA)$. *Show that the derivative of* z *with respect to the matrix* A *is*

$$\frac{\partial z}{\partial A} = CA + C^TA. \tag{36.3}$$

Therefore, if $C = Q$ *is symmetric, then the derivative is* $2QA$.

We have restricted the discussion here to real matrices and vectors. It often happens that we want to optimize a real quantity with respect to a complex vector. We can rewrite such quantities in terms of the real and imaginary parts of the complex values involved, to reduce everything to the real case just considered. For example, let Q be a hermitian matrix; then the quadratic form $\mathbf{k}^\dagger Q\mathbf{k}$ is real, for any complex vector $\mathbf{k}$. As we saw in Exercise 5.9, we can write the quadratic form entirely in terms of real matrices and vectors.

If $w = u + iv$ is a complex number with real part u and imaginary part v, the function $z = f(w) = |w|^2$ is real-valued. The derivative of $z = f(w)$ with respect to the complex variable w does not exist. When we write $z = u^2 + v^2$, we consider z as a function of the real vector $\mathbf{x} = (u, v)^T$. The derivative of z with respect to $\mathbf{x}$ is the vector $(2u, 2v)^T$.

Similarly, when we consider the real quadratic form $\mathbf{k}^\dagger Q\mathbf{k}$, we view each of the complex entries of the N by 1 vector $\mathbf{k}$ as two real numbers forming a two-dimensional real vector. We then differentiate the quadratic form with respect to the $2N$ by 1 real vector formed from these real and imaginary parts. If we turn the resulting $2N$ by 1 real vector back into an N by 1 complex vector, we get $2Q\mathbf{k}$ as the derivative; so, it appears as if the formula for differentiating in the real case carries over to the complex case.

36.5 Eigenvectors and Optimization

We can use these results concerning differentiation with respect to a vector to show that eigenvectors solve certain optimization problems.

Consider the problem of maximizing the quadratic form $\mathbf{x}^\dagger Q\mathbf{x}$, subject to $\mathbf{x}^\dagger\mathbf{x} = 1$; here the matrix Q is Hermitian, positive-definite, so that all of its eigenvalues are positive. We use the Lagrange-multiplier approach, with the Lagrangian

$$L(\mathbf{x}, \lambda) = \mathbf{x}^\dagger Q\mathbf{x} - \lambda\mathbf{x}^\dagger\mathbf{x},$$

where the scalar variable λ is the Lagrange multiplier. We differentiate $L(\mathbf{x}, \lambda)$ with respect to $\mathbf{x}$ and set the result equal to zero, obtaining

$$2Q\mathbf{x} - 2\lambda\mathbf{x} = 0,$$

or

$$Q\mathbf{x} = \lambda\mathbf{x}.$$

Therefore, $\mathbf{x}$ is an eigenvector of Q and λ is its eigenvalue. Since

$$\mathbf{x}^\dagger Q\mathbf{x} = \lambda\mathbf{x}^\dagger\mathbf{x} = \lambda,$$

we conclude that $\lambda = \lambda_1$, the largest eigenvalue of Q, and $\mathbf{x} = \mathbf{u}^1$, a norm-one eigenvector associated with λ_1.

Now consider the problem of maximizing $\mathbf{x}^\dagger Q\mathbf{x}$, subject to $\mathbf{x}^\dagger\mathbf{x} = 1$, and $\mathbf{x}^\dagger\mathbf{u}^1 = 0$. The Lagrangian is now

$$L(\mathbf{x}, \lambda, \alpha) = \mathbf{x}^\dagger Q\mathbf{x} - \lambda\mathbf{x}^\dagger\mathbf{x} - \alpha\mathbf{x}^\dagger\mathbf{u}^1.$$

Differentiating with respect to the vector $\mathbf{x}$ and setting the result equal to zero, we find that

$$2Q\mathbf{x} - 2\lambda\mathbf{x} - \alpha\mathbf{u}^1 = 0,$$

or

$$Q\mathbf{x} = \lambda\mathbf{x} + \beta\mathbf{u}^1,$$

for $\beta = \alpha/2$. But, we know that

$$(\mathbf{u}^1)^\dagger Q\mathbf{x} = \lambda(\mathbf{u}^1)^\dagger\mathbf{x} + \beta(\mathbf{u}^1)^\dagger\mathbf{u}^1 = \beta,$$

and

$$(\mathbf{u}^1)^\dagger Q\mathbf{x} = (Q\mathbf{u}^1)^\dagger\mathbf{x} = \lambda_1(\mathbf{u}^1)^\dagger\mathbf{x} = 0,$$

so $\beta = 0$ and we have

$$Q\mathbf{x} = \lambda\mathbf{x}.$$

Since

$$\mathbf{x}^\dagger Q\mathbf{x} = \lambda,$$

we conclude that $\mathbf{x}$ is a norm-one eigenvector of Q associated with the second-largest eigenvalue, $\lambda = \lambda_2$.

Continuing in this fashion, we can show that the norm-one eigenvector of Q associated with the nth largest eigenvalue λ_n maximizes the quadratic form $\mathbf{x}^\dagger Q\mathbf{x}$, subject to the constraints $\mathbf{x}^\dagger\mathbf{x} = 1$ and $\mathbf{x}^\dagger\mathbf{u}^m = 0$, for $m = 1, 2, ..., n-1$.

Chapter 37

Appendix: Signal Detection and Estimation

37.1 Chapter Summary

In this chapter we consider the problem of deciding whether or not a particular signal is present in the measured data; this is the *detection* problem. The underlying framework for the detection problem is optimal estimation and statistical hypothesis testing [143].

37.2 The Model of Signal in Additive Noise

The basic model used in detection is that of a signal in additive noise. The complex data vector is $\mathbf{x} = (x_1, x_2, ..., x_N)^T$. We assume that there are two possibilities:

Case 1: Noise only

$$x_n = z_n,\ n = 1, ..., N,$$

or

Case 2: Signal in noise

$$x_n = \gamma s_n + z_n,$$

where $\mathbf{z} = (z_1, z_2, ..., z_N)^T$ is a complex vector whose entries z_n are values of random variables that we call *noise*, about which we have only statistical information (that is to say, information about the average behavior), $\mathbf{s} = (s_1, s_2, ..., s_N)^T$ is a complex signal vector that we may known exactly, or at least for which we have a specific parametric model, and γ is a scalar that may be viewed either as deterministic or random (but unknown, in either case). Unless otherwise stated, we shall assume that γ is deterministic.

The *detection problem* is to decide which case we are in, based on some calculation performed on the data $\mathbf{x}$. Since Case 1 can be viewed as a special case of Case 2 in which the value of γ is zero, the detection problem is closely related to the problem of estimating γ, which we discussed in the chapter dealing with the best linear unbiased estimator, the BLUE.

We shall assume throughout that the entries of $\mathbf{z}$ correspond to random variables with means equal to zero. What the variances are and whether or not these random variables are mutually correlated will be discussed next. In all cases we shall assume that this information has been determined previously and is available to us in the form of the covariance matrix $Q = E(\mathbf{z}\mathbf{z}^\dagger)$ of the vector $\mathbf{z}$; the symbol E denotes expected value, so the entries of Q are the quantities $Q_{mn} = E(z_m \overline{z}_n)$. The diagonal entries of Q are $Q_{nn} = \sigma_n^2$, the variance of z_n.

Note that we have adopted the common practice of using the same symbols, z_n, when speaking about the random variables and about the specific values of these random variables that are present in our data. The context should make it clear to which we are referring.

In Case 2 we say that the *signal power* is equal to $|\gamma|^2 \frac{1}{N} \sum_{n=1}^{N} |s_n|^2 = \frac{1}{N}|\gamma|^2 \mathbf{s}^\dagger \mathbf{s}$ and the *noise power* is $\frac{1}{N} \sum_{n=1}^{N} \sigma_n^2 = \frac{1}{N} tr(Q)$, where $tr(Q)$ is the trace of the matrix Q, that is, the sum of its diagonal terms; therefore, the noise power is the average of the variances σ_n^2. The *input signal-to-noise ratio* (SNR_{in}) is the ratio of the signal power to that of the noise, prior to processing the data; that is,

$$\mathrm{SNR}_{\mathrm{in}} = \frac{1}{N}|\gamma|^2 \mathbf{s}^\dagger \mathbf{s} / \frac{1}{N} tr(Q) = |\gamma|^2 \mathbf{s}^\dagger \mathbf{s} / tr(Q).$$

37.3 Optimal Linear Filtering for Detection

In each case to be considered next, our detector will take the form of a linear estimate of γ; that is, we shall compute the estimate $\hat{\gamma}$ given by

$$\hat{\gamma} = \sum_{n=1}^{N} \overline{b}_n x_n = \mathbf{b}^\dagger \mathbf{x},$$

where $\mathbf{b} = (b_1, b_2, ..., b_N)^T$ is a vector to be determined. The objective is to use what we know about the situation to select the optimal $\mathbf{b}$, which will depend on $\mathbf{s}$ and Q.

For any given vector $\mathbf{b}$, the quantity

$$\hat{\gamma} = \mathbf{b}^\dagger \mathbf{x} = \gamma \mathbf{b}^\dagger \mathbf{s} + \mathbf{b}^\dagger \mathbf{z}$$

is a random variable whose mean value is equal to $\gamma \mathbf{b}^\dagger \mathbf{s}$ and whose variance is

$$var(\hat{\gamma}) = E(|\mathbf{b}^\dagger \mathbf{z}|^2) = E(\mathbf{b}^\dagger \mathbf{z}\mathbf{z}^\dagger \mathbf{b}) = \mathbf{b}^\dagger E(\mathbf{z}\mathbf{z}^\dagger)\mathbf{b} = \mathbf{b}^\dagger Q \mathbf{b}.$$

Therefore, the *output signal-to-noise ratio* (SNR_{out}) is defined as

$$\text{SNR}_{\text{out}} = |\gamma \mathbf{b}^\dagger \mathbf{s}|^2 / \mathbf{b}^\dagger Q \mathbf{b}.$$

The advantage we obtain from processing the data is called the *gain* associated with $\mathbf{b}$ and is defined to be the ratio of the SNR_{out} to SNR_{in}; that is,

$$\text{gain}(\mathbf{b}) = \frac{|\gamma \mathbf{b}^\dagger \mathbf{s}|^2 / (\mathbf{b}^\dagger Q \mathbf{b})}{|\gamma|^2 (\mathbf{s}^\dagger \mathbf{s}) / tr(Q)} = \frac{|\mathbf{b}^\dagger \mathbf{s}|^2 \, tr(Q)}{(\mathbf{b}^\dagger Q \mathbf{b})(\mathbf{s}^\dagger \mathbf{s})}.$$

The best $\mathbf{b}$ to use will be the one for which $\text{gain}(\mathbf{b})$ is the largest. So, ignoring the terms in the gain formula that do not involve $\mathbf{b}$, we see that the problem becomes *maximize* $\frac{|\mathbf{b}^\dagger \mathbf{s}|^2}{\mathbf{b}^\dagger Q \mathbf{b}}$, *for fixed signal vector* $\mathbf{s}$ *and fixed noise covariance matrix* Q.

The Cauchy inequality plays a major role in optimal filtering and detection:

Cauchy's inequality: For any vectors $\mathbf{a}$ and $\mathbf{b}$ we have

$$|\mathbf{a}^\dagger \mathbf{b}|^2 \leq (\mathbf{a}^\dagger \mathbf{a})(\mathbf{b}^\dagger \mathbf{b}),$$

with equality if and only if $\mathbf{a}$ is proportional to $\mathbf{b}$; that is, there is a scalar β such that $\mathbf{b} = \beta \mathbf{a}$.

Ex. 37.1 *Use Cauchy's inequality to show that, for any fixed vector* $\mathbf{a}$*, the choice* $\mathbf{b} = \beta\mathbf{a}$ *maximizes the quantity* $|\mathbf{b}^\dagger\mathbf{a}|^2/\mathbf{b}^\dagger\mathbf{b}$*, for any constant* β*.*

Ex. 37.2 *Use the definition of the covariance matrix* Q *to show that* Q *is Hermitian and that, for any vector* $\mathbf{y}$*,* $\mathbf{y}^\dagger Q\mathbf{y} \geq 0$*. Therefore,* Q *is a nonnegative definite matrix and, using its eigenvector decomposition, can be written as* $Q = CC^\dagger$*, for some invertible square matrix* C*.*

Ex. 37.3 *Consider now the problem of maximizing* $|\mathbf{b}^\dagger\mathbf{s}|^2/\mathbf{b}^\dagger Q\mathbf{b}$*. Using the two previous exercises, show that the solution is* $\mathbf{b} = \beta Q^{-1}\mathbf{s}$*, for some arbitrary constant* β*.*

We can now use the results of these exercises to continue our discussion. We choose the constant $\beta = 1/(\mathbf{s}^\dagger Q^{-1}\mathbf{s})$ so that the optimal $\mathbf{b}$ has $\mathbf{b}^\dagger\mathbf{s} = 1$; that is, the *optimal filter* $\mathbf{b}$ is

$$\mathbf{b} = (1/(\mathbf{s}^\dagger Q^{-1}\mathbf{s}))Q^{-1}\mathbf{s},$$

and the *optimal estimate* of γ is

$$\hat{\gamma} = \mathbf{b}^\dagger\mathbf{x} = (1/(\mathbf{s}^\dagger Q^{-1}\mathbf{s}))(\mathbf{s}^\dagger Q^{-1}\mathbf{x}).$$

The mean of the random variable $\hat{\gamma}$ is equal to $\gamma\mathbf{b}^\dagger\mathbf{s} = \gamma$, and the variance is equal to $1/(\mathbf{s}^\dagger Q^{-1}\mathbf{s})$. Therefore, the output signal power is $|\gamma|^2$, the output noise power is $1/(\mathbf{s}^\dagger Q^{-1}\mathbf{s})$, and so the *output signal-to-noise ratio* (SNR_{out}) is

$$\mathrm{SNR}_{\mathrm{out}} = |\gamma|^2(\mathbf{s}^\dagger Q^{-1}\mathbf{s}).$$

The gain associated with the optimal vector $\mathbf{b}$ is then

$$\text{maximum gain} = \frac{(\mathbf{s}^\dagger Q^{-1}\mathbf{s})\, tr(Q)}{(\mathbf{s}^\dagger\mathbf{s})}.$$

The calculation of the vector $C^{-1}\mathbf{x}$ is sometimes called *prewhitening* since $C^{-1}\mathbf{x} = \gamma C^{-1}\mathbf{s} + C^{-1}\mathbf{z}$ and the new noise vector, $C^{-1}\mathbf{z}$, has the identity matrix for its covariance matrix. The new signal vector is $C^{-1}\mathbf{s}$. The filtering operation that gives $\hat{\gamma} = \mathbf{b}^\dagger\mathbf{x}$ can be written as

$$\hat{\gamma} = (1/(\mathbf{s}^\dagger Q^{-1}\mathbf{s}))(C^{-1}\mathbf{s})^\dagger C^{-1}\mathbf{x};$$

the term $(C^{-1}\mathbf{s})^\dagger C^{-1}\mathbf{x}$ is described by saying that we *prewhiten, then do a matched filter*. Now we consider some special cases of noise.

37.4 The Case of White Noise

We say that the noise is *white noise* if the covariance matrix is $Q = \sigma^2 I$, where I denotes the identity matrix that is one on the main diagonal and zero elsewhere and $\sigma > 0$ is the common standard deviation of the z_n. This means that the z_n are mutually uncorrelated (independent, in the Gaussian case) and share a common variance.

In this case the optimal vector $\mathbf{b}$ is $\mathbf{b} = \frac{1}{(\mathbf{s}^\dagger \mathbf{s})}\mathbf{s}$ and the gain is N. Notice that $\hat{\gamma}$ now involves only a matched filter. We consider now some special cases of the signal vectors $\mathbf{s}$.

37.4.1 Constant Signal

Suppose that the vector $\mathbf{s}$ is constant; that is, $\mathbf{s} = \mathbf{1} = (1, 1, ..., 1)^T$. Then, we have

$$\hat{\gamma} = \frac{1}{N}\sum_{n=1}^{N} x_n.$$

This is the same result we found in our discussion of the BLUE, when we estimated the mean value and the noise was white.

37.4.2 Sinusoidal Signal, Frequency Known

Suppose that

$$\mathbf{s} = \mathbf{e}(\omega_0) = (\exp(-i\omega_0), \exp(-2i\omega_0), ..., \exp(-Ni\omega_0))^T,$$

where ω_0 denotes a known frequency in $[-\pi, \pi)$. Then, $\mathbf{b} = \frac{1}{N}\mathbf{e}(\omega_0)$ and

$$\hat{\gamma} = \frac{1}{N}\sum_{n=1}^{N} x_n \exp(in\omega_0);$$

so, we see yet another occurrence of the DFT.

37.4.3 Sinusoidal Signal, Frequency Unknown

If we do not know the value of the signal frequency ω_0, a reasonable thing to do is to calculate the $\hat{\gamma}$ for each (actually, finitely many) of the possible frequencies within $[-\pi, \pi)$ and base the detection decision on the largest value; that is, we calculate the DFT as a function of the variable ω. If there is only a single ω_0 for which there is a sinusoidal signal present in the data, the values of $\hat{\gamma}$ obtained at frequencies other than ω_0 provide estimates of the noise power σ^2, against which the value of $\hat{\gamma}$ for ω_0 can be compared.

37.5 The Case of Correlated Noise

We say that the noise is *correlated* if the covariance matrix Q is not a multiple of the identity matrix. This means either that the z_n are mutually correlated (dependent, in the Gaussian case) or that they are uncorrelated, but have different variances.

In this case, as we saw previously, the optimal vector $\mathbf{b}$ is

$$\mathbf{b} = \frac{1}{(\mathbf{s}^\dagger Q^{-1}\mathbf{s})} Q^{-1}\mathbf{s}$$

and the gain is

$$\text{maximum gain} = \frac{(\mathbf{s}^\dagger Q^{-1}\mathbf{s})\, tr(Q)}{(\mathbf{s}^\dagger \mathbf{s})}.$$

How large or small the gain is depends on how the signal vector $\mathbf{s}$ relates to the matrix Q.

For sinusoidal signals, the quantity $\mathbf{s}^\dagger\mathbf{s}$ is the same, for all values of the parameter ω; this is not always the case, however. In passive detection of sources in acoustic array processing, for example, the signal vectors arise from models of the acoustic medium involved. For far-field sources in an (acoustically) isotropic deep ocean, planewave models for $\mathbf{s}$ will have the property that $\mathbf{s}^\dagger\mathbf{s}$ does not change with source location. However, for near-field or shallow-water environments, this is usually no longer the case.

It follows from Exercise 37.3 that the quantity $\frac{\mathbf{s}^\dagger Q^{-1}\mathbf{s}}{\mathbf{s}^\dagger\mathbf{s}}$ achieves its maximum value when $\mathbf{s}$ is an eigenvector of Q associated with its smallest eigenvalue, λ_N; in this case, we are saying that the signal vector does not look very much like a typical noise vector. The maximum gain is then $\lambda_N^{-1} tr(Q)$. Since $tr(Q)$ equals the sum of its eigenvalues, multiplying by $tr(Q)$ serves to normalize the gain, so that we cannot get larger gain simply by having all the eigenvalues of Q small.

On the other hand, if $\mathbf{s}$ should be an eigenvector of Q associated with its largest eigenvalue, say λ_1, then the maximum gain is $\lambda_1^{-1} tr(Q)$. If the noise is signal-like, that is, has one dominant eigenvalue, then $tr(Q)$ is approximately λ_1 and the maximum gain is around one, so we have lost the maximum gain of N we were able to get in the white-noise case. This makes sense, in that it says that we cannot significantly improve our ability to discriminate between signal and noise by taking more samples, if the signal and noise are very similar.

37.5.1 Constant Signal with Unequal-Variance Uncorrelated Noise

Suppose that the vector $\mathbf{s}$ is constant; that is, $\mathbf{s} = \mathbf{1} = (1, 1, ..., 1)^T$. Suppose also that the noise covariance matrix is $Q = \text{diag}\{\sigma_1, ..., \sigma_N\}$.

In this case the optimal vector $\mathbf{b}$ has entries

$$b_m = \frac{1}{(\sum_{n=1}^{N} \sigma_n^{-1})} \sigma_m^{-1},$$

for $m = 1, ..., N$, and we have

$$\hat{\gamma} = \frac{1}{(\sum_{n=1}^{N} \sigma_n^{-1})} \sum_{m=1}^{N} \sigma_m^{-1} x_m.$$

This is the BLUE estimate of γ in this case.

37.5.2 Sinusoidal signal, Frequency Known, in Correlated Noise

Suppose that

$$\mathbf{s} = \mathbf{e}(\omega_0) = (\exp(-i\omega_0), \exp(-2i\omega_0), ..., \exp(-Ni\omega_0))^T,$$

where ω_0 denotes a known frequency in $[-\pi, \pi)$. In this case the optimal vector $\mathbf{b}$ is

$$\mathbf{b} = \frac{1}{\mathbf{e}(\omega_0)^\dagger Q^{-1} \mathbf{e}(\omega_0)} Q^{-1} \mathbf{e}(\omega_0)$$

and the gain is

$$\text{maximum gain} = \frac{1}{N} [\mathbf{e}(\omega_0)^\dagger Q^{-1} \mathbf{e}(\omega_0)] tr(Q).$$

How large or small the gain is depends on the quantity $q(\omega_0)$, where

$$q(\omega) = \mathbf{e}(\omega)^\dagger Q^{-1} \mathbf{e}(\omega).$$

The function $1/q(\omega)$ can be viewed as a sort of noise power spectrum, describing how the noise power appears when decomposed over the various frequencies in $[-\pi, \pi)$. The maximum gain will be large if this *noise power spectrum* is relatively small near $\omega = \omega_0$; however, when the noise is similar to the signal, that is, when the noise power spectrum is relatively large near $\omega = \omega_0$, the maximum gain can be small. In this case the noise power spectrum plays a role analogous to that played by the eigenvalues of Q earlier.

To see more clearly why it is that the function $1/q(\omega)$ can be viewed

as a sort of noise power spectrum, consider what we get when we apply the optimal filter associated with ω to data containing only noise. The average output should tell us how much power there is in the component of the noise that resembles $\mathbf{e}(\omega)$; this is essentially what is meant by a noise power spectrum. The result is $\mathbf{b}^\dagger\mathbf{z} = (1/q(\omega))\mathbf{e}(\omega)^\dagger Q^{-1}\mathbf{z}$. The expected value of $|\mathbf{b}^\dagger\mathbf{z}|^2$ is then $1/q(\omega)$.

37.5.3 Sinusoidal Signal, Frequency Unknown, in Correlated Noise

Again, if we do not know the value of the signal frequency ω_0, a reasonable thing to do is to calculate the $\hat{\gamma}$ for each (actually, finitely many) of the possible frequencies within $[-\pi, \pi)$ and base the detection decision on the largest value. For each ω the corresponding value of $\hat{\gamma}$ is

$$\hat{\gamma}(\omega) = [1/(\mathbf{e}(\omega)^\dagger Q^{-1}\mathbf{e}(\omega))] \sum_{n=1}^{N} a_n \exp(in\omega),$$

where $\mathbf{a} = (a_1, a_2, ..., a_N)^T$ satisfies the linear system $Q\mathbf{a} = \mathbf{x}$ or $\mathbf{a} = Q^{-1}\mathbf{x}$. It is interesting to note the similarity between this estimation procedure and the PDFT discussed earlier; to see the connection, view $[1/(\mathbf{e}(\omega)^\dagger Q^{-1}\mathbf{e}(\omega))]$ in the role of $P(\omega)$ and Q its corresponding matrix of Fourier-transform values. The analogy breaks down when we notice that Q need not be Toeplitz, as in the PDFT case; however, the similarity is intriguing.

37.6 Capon's Data-Adaptive Method

When the noise covariance matrix Q is not available, perhaps because we cannot observe the background noise in the absence of any signals that may also be present, we may use the signal-plus-noise covariance matrix R in place of Q.

Ex. 37.4 *Show that for*

$$R = |\gamma|^2 ss^\dagger + Q$$

maximizing the ratio

$$|b^\dagger s|^2 / b^\dagger R b$$

is equivalent to maximizing the ratio

$$|b^\dagger s|^2 / b^\dagger Q b.$$

In [76] Capon offered a high-resolution method for detecting and resolving sinusoidal signals with unknown frequencies in noise. His estimator has the form

$$1/e(\omega)^\dagger R^{-1} e(\omega). \tag{37.1}$$

The idea here is to fix an arbitrary ω, and then to find the vector $b(\omega)$ that minimizes $b(\omega)^\dagger R b(\omega)$, subject to $b(\omega)^\dagger e(\omega) = 1$. The vector $b(\omega)$ turns out to be

$$b(\omega) = \frac{1}{e(\omega)^\dagger R^{-1} e(\omega)} R^{-1} e(\omega). \tag{37.2}$$

Now we allow ω to vary and compute the expected output of the filter $b(\omega)$, operating on the signal plus noise input. This expected output is then

$$1/e(\omega)^\dagger R^{-1} e(\omega). \tag{37.3}$$

The reason that this estimator resolves closely spaced delta functions better than linear methods such as the DFT is that, when ω is fixed, we obtain an optimal filter using R as the noise covariance matrix, which then includes all sinusoids not at the frequency ω in the noise component. This is actually a good thing, since, when we are looking at a frequency ω that does not correspond to a frequency actually present in the data, we want the sinusoidal components present at nearby frequencies to be filtered out.

Bibliography

[1] Agmon, S. (1954) "The relaxation method for linear inequalities." *Canadian Journal of Mathematics* **6**, pp. 382–392.

[2] Ahn, S., and Fessler, J. (2003) "Globally convergent image reconstruction for emission tomography using relaxed ordered subset algorithms." *IEEE Transactions on Medical Imaging*, **22(5)**, pp. 613–626.

[3] Ahn, S., Fessler, J., Blatt, D., and Hero, A. (2006) "Convergent incremental optimization transfer algorithms: application to tomography." *IEEE Transactions on Medical Imaging*, **25(3)**, pp. 283–296.

[4] Anderson, T. (1972) "Efficient estimation of regression coefficients in time series." *Proc. of Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: The Theory of Statistics* University of California Press, Berkeley, CA, pp. 471–482.

[5] Anderson, A. and Kak, A. (1984) "Simultaneous algebraic reconstruction technique (SART): a superior implementation of the ART algorithm." *Ultrasonic Imaging* **6**, pp. 81–94.

[6] Ash, R. and Gardner, M. (1975) *Topics in Stochastic Processes* Boston: Academic Press.

[7] Axelsson, O. (1994) *Iterative Solution Methods.* Cambridge, UK: Cambridge University Press.

[8] Baillet, S., Mosher, J., and Leahy, R. (2001) "Electromagnetic Brain Mapping" , *IEEE Signal Processing Magazine*, **18 (6)**, pp. 14–30.

[9] Baillon, J.-B., Bruck, R.E., and Reich, S. (1978) "On the asymptotic behavior of nonexpansive mappings and semigroups in Banach spaces" , *Houston Journal of Mathematics*, **4**, pp. 1–9.

[10] Barrett, H., White, T., and Parra, L. (1997) "List-mode likelihood." *J. Opt. Soc. Am. A* **14**, pp. 2914–2923.

[11] Bauschke, H. (1996) "The approximation of fixed points of compositions of nonexpansive mappings in Hilbert space," *Journal of Mathematical Analysis and Applications*, **202**, pp. 150–159.

[12] Bauschke, H. (2001) "Projection algorithms: results and open problems."in *Inherently Parallel Algorithms in Feasibility and Optimization and their Applications*, Butnariu, D., Censor, Y., and Reich, S., editors, Amsterdam: Elsevier Science. pp. 11–22.

[13] Bauschke, H. and Borwein, J. (1996) "On projection algorithms for solving convex feasibility problems."*SIAM Review* **38 (3)**, pp. 367–426.

[14] Bauschke, H., Borwein, J., and Lewis, A. (1997) "The method of cyclic projections for closed convex sets in Hilbert space."*Contemporary Mathematics: Recent Developments in Optimization Theory and Nonlinear Analysis* **204**, American Mathematical Society, pp. 1–38.

[15] Bauschke, H., and Lewis, A. (2000) "Dykstra's algorithm with Bregman projections: a convergence proof." *Optimization*, **48**, pp. 409–427.

[16] Benson, M. (2003) "What *Galileo* Saw." in *The New Yorker*; reprinted in [95].

[17] Bertero, M. (1992) "Sampling theory, resolution limits and inversion methods."in [19], pp. 71–94.

[18] Bertero, M., and Boccacci, P. (1998) *Introduction to Inverse Problems in Imaging* Bristol, UK: Institute of Physics Publishing.

[19] Bertero, M. and Pike, E.R., editors (1992) *Inverse Problems in Scattering and Imaging* Malvern Physics Series, Adam Hilger, IOP Publishing, London.

[20] Bertsekas, D.P. (1997) "A new class of incremental gradient methods for least squares problems."*SIAM J. Optim.* **7**, pp. 913–926.

[21] Blackman, R. and Tukey, J. (1959) *The Measurement of Power Spectra*. New York: Dover Publications.

[22] Boas, D., Brooks, D., Miller, E., DiMarzio, C., Kilmer, M., Gaudette, R., and Zhang, Q. (2001) "Imaging the body with diffuse optical tomography." *IEEE Signal Processing Magazine*, **18 (6)**, pp. 57–75.

[23] Bochner, S. and Chandrasekharan, K. (1949) *Fourier Transforms*, Annals of Mathematical Studies, No. 19. Princeton, NJ: Princeton University Press.

[24] Born, M. and Wolf, E. (1999) *Principles of Optics: 7th edition*. Cambridge, UK: Cambridge University Press.

[25] Bouten, L., van Handel, R., and James, M. ((2009) "A discrete invitation to quantum filtering and feedback control." *SIAM Review*, **51(2)**, pp. 239–316.

[26] Borwein, J. and Lewis, A. (2000) *Convex Analysis and Nonlinear Optimization*. Canadian Mathematical Society Books in Mathematics, New York: Springer-Verlag.

[27] Bracewell, R.C. (1979) "Image reconstruction in radio astronomy." in [162], pp. 81–104.

[28] Brauer, A. (1946) "Characteristic roots of a matrix." *Duke Mathematics Journal*, **13**, pp. 387–395.

[29] Bregman, L.M. (1967) "The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming." *USSR Computational Mathematics and Mathematical Physics* **7**, pp. 200–217.

[30] Bregman, L., Censor, Y., and Reich, S. (1999) "Dykstra's algorithm as the nonlinear extension of Bregman's optimization method." *Journal of Convex Analysis*, **6 (2)**, pp. 319–333.

[31] Brooks, D., and MacLeod, R. (1997) "Electrical imaging of the heart." *IEEE Signal Processing Magazine*, **14 (1)**, pp. 24–42.

[32] Browne, E. (1930) "The characteristic roots of a matrix." *Bulletin of the American Mathematical Society*, **36**, pp. 705–710.

[33] Browne, J. and A. DePierro, A. (1996) "A row-action alternative to the EM algorithm for maximizing likelihoods in emission tomography." *IEEE Trans. Med. Imag.* **15**, pp. 687–699.

[34] Bruck, R.E., and Reich, S. (1977) "Nonexpansive projections and resolvents of accretive operators in Banach spaces." *Houston Journal of Mathematics*, **3**, pp. 459–470.

[35] Bruckstein, A., Donoho, D., and Elad, M. (2009) "From sparse solutions of systems of equations to sparse modeling of signals and images." *SIAM Review*, **51(1)**, pp. 34–81.

[36] Bruyant, P., Sau, J., and Mallet, J.J. (1999) "Noise removal using factor analysis of dynamic structures: application to cardiac gated studies." *Journal of Nuclear Medicine* **40 (10)**, pp. 1676–1682.

[37] Budinger, T., Gullberg, G., and Huesman, R. (1979) "Emission computed tomography." in [162], pp. 147–246.

[38] Burg, J. (1967) "Maximum entropy spectral analysis." *paper presented at the 37th Annual SEG meeting, Oklahoma City, OK.*

[39] Burg, J. (1972) "The relationship between maximum entropy spectra and maximum likelihood spectra." *Geophysics* **37**, pp. 375–376.

[40] Burg, J. (1975) *Maximum Entropy Spectral Analysis*, Ph.D. dissertation, Stanford University.

[41] Byrne, C. and Fitzgerald, R. (1979) "A unifying model for spectrum estimation." in *Proceedings of the RADC Workshop on Spectrum Estimation- October 1979*, Griffiss AFB, Rome, NY.

[42] Byrne, C. and Fitzgerald, R. (1982) "Reconstruction from partial information, with applications to tomography." *SIAM J. Applied Math.* **42(4)**, pp. 933–940.

[43] Byrne, C., Fitzgerald, R., Fiddy, M., Hall, T. and Darling, A. (1983) "Image restoration and resolution enhancement." *J. Opt. Soc. Amer.* **73**, pp. 1481–1487.

[44] Byrne, C., and Wells, D. (1983) "Limit of continuous and discrete finite-band Gerchberg iterative spectrum extrapolation." *Optics Letters* **8 (10)**, pp. 526–527.

[45] Byrne, C. and Fitzgerald, R. (1984) "Spectral estimators that extend the maximum entropy and maximum likelihood methods." *SIAM J. Applied Math.* **44(2)**, pp. 425–442.

[46] Byrne, C., Levine, B.M., and Dainty, J.C. (1984) "Stable estimation of the probability density function of intensity from photon frequency counts." *JOSA Communications* **1(11)**, pp. 1132–1135.

[47] Byrne, C., and Wells, D. (1985) "Optimality of certain iterative and non-iterative data extrapolation procedures." *Journal of Mathematical Analysis and Applications* **111 (1)**, pp. 26–34.

[48] Byrne, C. and Fiddy, M. (1987) "Estimation of continuous object distributions from Fourier magnitude measurements." *JOSA A* **4**, pp. 412–417.

[49] Byrne, C. and Fiddy, M. (1988) "Images as power spectra; reconstruction as Wiener filter approximation." *Inverse Problems* **4**, pp. 399–409.

[50] Byrne, C., Haughton, D., and Jiang, T. (1993) "High-resolution inversion of the discrete Poisson and binomial transformations." *Inverse Problems* **9**, pp. 39–56.

[51] Byrne, C. (1993) "Iterative image reconstruction algorithms based on cross-entropy minimization." *IEEE Transactions on Image Processing* **IP-2**, pp. 96–103.

[52] Byrne, C. (1995) "Erratum and addendum to 'Iterative image reconstruction algorithms based on cross-entropy minimization'." *IEEE Transactions on Image Processing* **IP-4**, pp. 225–226.

[53] Byrne, C. (1996) "Iterative reconstruction algorithms based on cross-entropy minimization."in *Image Models (and their Speech Model Cousins)*, S.E. Levinson and L. Shepp, editors, IMA Volumes in Mathematics and its Applications, Volume 80, pp. 1–11. New York: Springer-Verlag.

[54] Byrne, C. (1996) "Block-iterative methods for image reconstruction from projections." *IEEE Transactions on Image Processing* **IP-5**, pp. 792–794.

[55] Byrne, C. (1997) "Convergent block-iterative algorithms for image reconstruction from inconsistent data." *IEEE Transactions on Image Processing* **IP-6**, pp. 1296–1304.

[56] Byrne, C. (1998) "Accelerating the EMML algorithm and related iterative algorithms by rescaled block-iterative (RBI) methods." *IEEE Transactions on Image Processing* **IP-7**, pp. 100–109.

[57] Byrne, C. (1998) "Iterative deconvolution and deblurring with constraints." *Inverse Problems*, **14**, pp. 1455-1467.

[58] Byrne, C. (1999) "Iterative projection onto convex sets using multiple Bregman distances." *Inverse Problems* **15**, pp. 1295–1313.

[59] Byrne, C. (2000) "Block-iterative interior point optimization methods for image reconstruction from limited data." *Inverse Problems* **16**, pp. 1405–1419.

[60] Byrne, C. (2001) "Bregman-Legendre multidistance projection algorithms for convex feasibility and optimization."in *Inherently Parallel Algorithms in Feasibility and Optimization and their Applications*, Butnariu, D., Censor, Y., and Reich, S., editors, pp. 87–100. Amsterdam: Elsevier Publ.,

[61] Byrne, C. (2001) "Likelihood maximization for list-mode emission tomographic image reconstruction." *IEEE Transactions on Medical Imaging* **20(10)**, pp. 1084–1092.

[62] Byrne, C. (2002) "Iterative oblique projection onto convex sets and the split feasibility problem." *Inverse Problems* **18**, pp. 441–453.

[63] Byrne, C. (2004) "A unified treatment of some iterative algorithms in signal processing and image reconstruction." *Inverse Problems* **20**, pp. 103–120.

[64] Byrne, C. (2005) "Choosing parameters in block-iterative or ordered-subset reconstruction algorithms." *IEEE Transactions on Image Processing*, **14 (3)**, pp. 321–327.

[65] Byrne, C. (2005) *Signal Processing: A Mathematical Approach*, AK Peters, Publ., Wellesley, MA.

[66] Byrne, C. (2007) *Applied Iterative Methods*, AK Peters, Publ., Wellesley, MA.

[67] Byrne, C. (2008) "Sequential unconstrained minimization algorithms for constrained optimization." *Inverse Problems*, **24(1)**, article no. 015013.

[68] Byrne, C. (2009) "Block-iterative algorithms." *International Transactions in Operations Research*, **16(4)**.

[69] Byrne, C. (2009) "Bounds on the largest singular value of a matrix and the convergence of simultaneous and block-iterative algorithms for sparse linear systems." *International Transactions in Operations Research*, **16(4)**.

[70] Byrne, C. (2009) *A First Course in Optimization*, unpublished text available at my website.

[71] Byrne, C. and Censor, Y. (2001) "Proximity function minimization using multiple Bregman projections, with applications to split feasibility and Kullback-Leibler distance minimization." *Annals of Operations Research* **105**, pp. 77–98.

[72] Candès, E., and Romberg, J. (2007) "Sparsity and incoherence in compressive sampling." *Inverse Problems*, **23(3)**, pp. 969–985.

[73] Candès, E., Romberg, J., and Tao, T. (2006) "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information." *IEEE Transactions on Information Theory*, **52(2)**, pp. 489–509.

[74] Candès, E., Wakin, M., and Boyd, S. (2007) "Enhancing sparsity by reweighted l_1 minimization." preprint available at http://www.acm.caltech.edu/ emmanuel/publications.html .

[75] Candy, J. (1988) *Signal Processing: The Modern Approach* New York: McGraw-Hill Publ.

[76] Capon, J. (1969) "High-resolution frequency-wavenumber spectrum analysis." *Proc. of the IEEE* **57**, pp. 1408–1418.

[77] Carlson, D., Johnson, C., Lay, D., and Porter, A.D. (2002) *Linear Algebra Gems: Assets for Undergraduates*, The Mathematical Society of America, MAA Notes **59**.

[78] Cederquist, J., Fienup, J., Wackerman, C., Robinson, S., and Kryskowski, D. (1989) "Wave-front phase estimation from Fourier intensity measurements." *Journal of the Optical Society of America A* **6(7)**, pp. 1020–1026.

[79] Censor, Y. (1981) "Row-action methods for huge and sparse systems and their applications." *SIAM Review*, **23**: 444–464.

[80] Censor, Y., Eggermont, P.P.B., and Gordon, D. (1983) "Strong underrelaxation in Kaczmarz's method for inconsistent systems." *Numerische Mathematik* **41**, pp. 83–92.

[81] Censor, Y. and Elfving, T. (1994) "A multi-projection algorithm using Bregman projections in a product space." *Numerical Algorithms*, **8**, pp. 221–239.

[82] Censor, Y., Bortfeld, T., Martin, B., and Trofimov, A. "A unified approach for inversion problems in intensity-modulated radiation therapy." *Physics in Medicine and Biology* 51 (2006), pp. 2353-2365.

[83] Censor, Y., Elfving, T., Herman, G.T., and Nikazad, T. (2008) "On diagonally-relaxed orthogonal projection methods." *SIAM Journal on Scientific Computation*, **30(1)**, pp. 473–504.

[84] Censor, Y., Elfving, T., Kopf, N., and Bortfeld, T. "The multiple-sets split feasibility problem and its application for inverse problems." *Inverse Problems* 21 (2005), pp. 2071-2084.

[85] Censor, Y., Gordon, D., and Gordon, R. (2001) "Component averaging: an efficient iterative parallel algorithm for large and sparse unstructured problems." *Parallel Computing*, **27**, pp. 777–808.

[86] Censor, Y., Gordon, D., and Gordon, R. (2001) "BICAV: A block-iterative, parallel algorithm for sparse systems with pixel-related weighting." *IEEE Transactions on Medical Imaging*, **20**, pp. 1050–1060.

[87] Censor, Y., and Reich, S. (1996) "Iterations of paracontractions and firmly nonexpansive operators with applications to feasibility and optimization" , *Optimization*, **37**, pp. 323–339.

[88] Censor, Y., and Reich, S. (1998) "The Dykstra algorithm for Bregman projections." *Communications in Applied Analysis*, **2**, pp. 323–339.

[89] Censor, Y. and Segman, J. (1987) "On block-iterative maximization."*J. of Information and Optimization Sciences* **8**, pp. 275–291.

[90] Censor, Y. and Zenios, S.A. (1997) *Parallel Optimization: Theory, Algorithms and Applications*. New York: Oxford University Press.

[91] Chang, J.-H., Anderson, J.M.M., and Votaw, J.R. (2004) "Regularized image reconstruction algorithms for positron emission tomography." *IEEE Transactions on Medical Imaging* **23(9)**, pp. 1165–1175.

[92] Childers, D., editor (1978) *Modern Spectral Analysis*. New York:IEEE Press.

[93] Chui, C. and Chen, G. (1991) *Kalman Filtering*, second edition. Berlin: Springer-Verlag.

[94] Cimmino, G. (1938) "Calcolo approssimato per soluzioni dei sistemi di equazioni lineari." *La Ricerca Scientifica XVI, Series II, Anno IX* **1**, pp. 326–333.

[95] Cohen, J. (2010) (editor) *The Best of The Best American Science Writing*, Harper-Collins Publ.

[96] Combettes, P. (1993) "The foundations of set theoretic estimation." *Proceedings of the IEEE* **81 (2)**, pp. 182–208.

[97] Combettes, P. (1996) "The convex feasibility problem in image recovery." *Advances in Imaging and Electron Physics* **95**, pp. 155–270.

[98] Combettes, P. (2000) "Fejér monotonicity in convex optimization." in *Encyclopedia of Optimization*, C.A. Floudas and P. M. Pardalos, editors, Boston: Kluwer Publ.

[99] Combettes, P., and Trussell, J. (1990) "Method of successive projections for finding a common point of sets in a metric space." *Journal of Optimization Theory and Applications* **67 (3)**, pp. 487–507.

[100] Combettes, P., and Wajs, V. (2005) "Signal recovery by proximal forward-backward splitting." *Multi-scale Modeling and Simulation*, **4(4)**, pp. 1168–1200.

[101] Cooley, J. and Tukey, J. (1965) "An algorithm for the machine calculation of complex Fourier series." *Math. Comp.*, **19**, pp. 297–301.

[102] Csiszár, I. (1989) "A geometric interpretation of Darroch and Ratcliff's generalized iterative scaling." *The Annals of Statistics* **17 (3)**, pp. 1409–1413.

[103] Csiszár, I. (1991) "Why least squares and maximum entropy? An axiomatic approach to inference for linear inverse problems." *The Annals of Statistics* **19 (4)**, pp. 2032–2066.

[104] Csiszár, I. and Tusnády, G. (1984) "Information geometry and alternating minimization procedures." *Statistics and Decisions* **Supp. 1**, pp. 205–237.

[105] Cullen, C. (1966) *Matrices and Linear Transformations*. Reading, MA: Addison-Wesley.

[106] Dainty, J. C. and Fiddy, M. (1984) "The essential role of prior knowledge in phase retrieval." *Optica Acta* **31**, pp. 325–330.

[107] Darroch, J. and Ratcliff, D. (1972) "Generalized iterative scaling for log-linear models." *Annals of Mathematical Statistics* **43**, pp. 1470–1480.

[108] Dax, A. (1990) "The convergence of linear stationary iterative processes for solving singular unstructured systems of linear equations." *SIAM Review*, **32**, pp. 611–635.

[109] Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977) "Maximum likelihood from incomplete data via the EM algorithm." *Journal of the Royal Statistical Society, Series B* **37**, pp. 1–38.

[110] De Pierro, A. (1995) "A modified expectation maximization algorithm for penalized likelihood estimation in emission tomography." *IEEE Transactions on Medical Imaging* **14**, pp. 132–137.

[111] De Pierro, A. and Iusem, A. (1990) "On the asymptotic behavior of some alternate smoothing series expansion iterative methods." *Linear Algebra and its Applications* **130**, pp. 3–24.

[112] De Pierro, A., and Yamaguchi, M. (2001) "Fast EM-like methods for maximum 'a posteriori' estimates in emission tomography." *Transactions on Medical Imaging*, **20 (4)**.

[113] Deutsch, F., and Yamada, I. (1998) "Minimizing certain convex functions over the intersection of the fixed point sets of nonexpansive mappings." *Numerical Functional Analysis and Optimization*, **19**, pp. 33–56.

[114] Dhanantwari, A., Stergiopoulos, S., and Iakovidis, I. (2001) "Correcting organ motion artifacts in x-ray CT medical imaging systems by adaptive processing. I. Theory." *Med. Phys.* **28(8)**, pp. 1562–1576.

[115] Dines, K., and Lyttle, R. (1979) "Computerized geophysical tomography." *Proc. IEEE*, **67**, pp. 1065–1073.

[116] Donoho, D. (2006) "Compressed sampling." *IEEE Transactions on Information Theory*, **52 (4)**. (download preprints at http://www.stat.stanford.edu/ donoho/Reports).

[117] Driscoll, P., and Fox, W. (1996) "Presenting the Kuhn-Tucker conditions using a geometric method." *The College Mathematics Journal*, **38 (1)**, pp. 101–108.

[118] Drmač, Z., and Veselić, K. (2008) "New fast and accurate Jacobi SVD algorithms: Part I. " *SIAM J. Matrix Anal. Appl.*, **29**, pp. 1322–1342.

[119] Drmač, Z., and Veselić, K. (2008) "New fast and accurate Jacobi SVD algorithms: Part II. " *SIAM J. Matrix Anal. Appl.*, **29**, pp. 1343–1362.

[120] Duda, R., Hart, P., and Stork, D. (2001) *Pattern Classification*, Wiley.

[121] Duffin, R., Peterson, E., and Zener, C. (1967) *Geometric Programming: Theory and Applications*. New York: Wiley.

[122] Dugundji, J. (1970) *Topology* Boston: Allyn and Bacon, Inc.

[123] Dykstra, R. (1983) "An algorithm for restricted least squares regression." *J. Amer. Statist. Assoc.*, **78 (384)**, pp. 837–842.

[124] Eggermont, P.P.B., Herman, G.T., and Lent, A. (1981) "Iterative algorithms for large partitioned linear systems, with applications to image reconstruction." *Linear Algebra and its Applications* **40**, pp. 37–67.

[125] Elsner, L., Koltracht, L., and Neumann, M. (1992) "Convergence of sequential and asynchronous nonlinear paracontractions." *Numerische Mathematik*, **62**, pp. 305–319.

[126] Erdogan, H., and Fessler, J. (1999) "Fast monotonic algorithms for transmission tomography." *IEEE Transactions on Medical Imaging*, **18(9)**, pp. 801–814.

[127] Everitt, B. and Hand, D. (1981) *Finite Mixture Distributions* London: Chapman and Hall.

[128] Farkas, J. (1902) "Über die Theorie der einfachen Ungleichungen." *J. Reine Angew. Math.*, **124**, pp. 1–24.

[129] Farncombe, T. (2000) "Functional dynamic SPECT imaging using a single slow camera rotation." *Ph.D. thesis, Dept. of Physics, University of British Columbia.*

[130] Farnell, A.B. (1944) "Limits for the characteristic roots of a matrix." *Bulletin of the American Mathematical Society*, **50**, pp. 789–794.

[131] Fernandez, J., Sorzano, C., Marabini, R., and Carazo, J-M. (2006) "Image processing and 3-D reconstruction in electron microscopy." *IEEE Signal Processing Magazine*, **23 (3)**, pp. 84–94.

[132] Fessler, J., Ficaro, E., Clinthorne, N., and Lange, K. (1997) "Grouped-coordinate ascent algorithms for penalized-likelihood transmission image reconstruction." *IEEE Transactions on Medical Imaging*, **16 (2)**, pp. 166–175.

[133] Feynman, R., Leighton, R., and Sands, M. (1963) *The Feynman Lectures on Physics, Vol. 1*. Boston: Addison-Wesley.

[134] Fiddy, M. (1983) "The phase retrieval problem." in *Inverse Optics*, SPIE Proceedings 413 (A.J. Devaney, editor), pp. 176–181.

[135] Fiddy, M. (2008) *private communication*.

[136] Fiacco, A., and McCormick, G. (1990) *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*. Philadelphia, PA: SIAM Classics in Mathematics (reissue).

[137] Fienup, J. (1979) "Space object imaging through the turbulent atmosphere." *Optical Engineering* **18**, pp. 529–534.

[138] Fienup, J. (1987) "Reconstruction of a complex-valued object from the modulus of its Fourier transform using a support constraint." *Journal of the Optical Society of America A* **4(1)**, pp. 118–123.

[139] Fleming, W. (1965) *Functions of Several Variables*, Addison-Wesley Publ., Reading, MA.

[140] Frieden, B. R. (1982) *Probability, Statistical Optics and Data Testing*. Berlin: Springer-Verlag.

[141] Gale, D. (1960) *The Theory of Linear Economic Models*. New York: McGraw-Hill.

[142] Gasquet, C. and Witomski, F. (1998) *Fourier Analysis and Applications*. Berlin: Springer-Verlag.

[143] Gelb, A., editor, (1974) *Applied Optimal Estimation*, written by the technical staff of The Analytic Sciences Corporation, MIT Press, Cambridge, MA.

[144] Geman, S., and Geman, D. (1984) "Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images." *IEEE Transactions on Pattern Analysis and Machine Intelligence* **PAMI-6**, pp. 721–741.

[145] Gerchberg, R. W. (1974) "Super-restoration through error energy reduction." *Optica Acta* **21**, pp. 709–720.

[146] Gifford, H., King, M., de Vries, D., and Soares, E. (2000) "Channelized Hotelling and human observer correlation for lesion detection in hepatic SPECT imaging." *Journal of Nuclear Medicine* **41(3)**, pp. 514–521.

[147] Goebel, K., and Reich, S. (1984) *Uniform Convexity, Hyperbolic Geometry, and Nonexpansive Mappings*, New York: Dekker.

[148] Golshtein, E., and Tretyakov, N. (1996) *Modified Lagrangians and Monotone Maps in Optimization*. New York: John Wiley and Sons, Inc.

[149] Golub, G., and Kahan, W. (1965) "Calculating the singular values and pseudo-inverse of a matrix." *SIAM J. Numer. Anal.*, Ser. B, **2**, pp. 205–224.

[150] Gordan, P. (1873) "Über die Auflösungen linearer Gleichungen mit reelen Coefficienten." *Math. Ann.*, **6**, pp. 23–28.

[151] Gordon, R., Bender, R., and Herman, G.T. (1970) "Algebraic reconstruction techniques (ART) for three-dimensional electron microscopy and x-ray photography." *J. Theoret. Biol.* **29**, pp. 471–481.

[152] Gordon, D., and Gordon, R.(2005) "Component-averaged row projections: A robust block-parallel scheme for sparse linear systems." *SIAM Journal on Scientific Computing*, **27**, pp. 1092–1117.

[153] Grcar, J. (2011) "John von Neumann's analysis of Gaussian elimination and the origins of modern numerical analysis." *SIAM Review*, **53 (4)**, pp. 607–682.

[154] Green, P. (1990) "Bayesian reconstructions from emission tomography data using a modified EM algorithm." *IEEE Transactions on Medical Imaging* **9**, pp. 84–93.

[155] Gubin, L.G., Polyak, B.T. and Raik, E.V. (1967) "The method of projections for finding the common point of convex sets." *USSR Computational Mathematics and Mathematical Physics* **7**, pp. 1–24.

[156] Gullberg, G., Huesman, R., Malko, J., Pelc, N., and Budinger, T. (1986) "An attenuated projector-backprojector for iterative SPECT reconstruction." *Physics in Medicine and Biology*, **30**, pp. 799–816.

[157] Haacke, E., Brown, R., Thompson, M., and Venkatesan, R. (1999) *Magnetic Resonance Imaging*. New York: Wiley-Liss.

[158] Hager, W. (1988) *Applied Numerical Linear Algebra*, Englewood Cliffs, NJ: Prentice- Hall.

[159] Hager, B., Clayton, R., Richards, M., Comer, R., and Dziewonsky, A. (1985) "Lower mantle heterogeneity, dynamic typography and the geoid." *Nature*, **313**, pp. 541–545.

[160] Haykin, S. (1985) *Array Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall.

[161] Hebert, T. and Leahy, R. (1989) "A generalized EM algorithm for 3-D Bayesian reconstruction from Poisson data using Gibbs priors." *IEEE Transactions on Medical Imaging* **8**, pp. 194–202.

[162] Herman, G.T. (ed.) (1979) *Image Reconstruction from Projections*, Topics in Applied Physics, Vol. 32, Springer-Verlag, Berlin.

[163] Herman, G.T., and Natterer, F. (eds.) (1981) *Mathematical Aspects of Computerized Tomography*, Lecture Notes in Medical Informatics, Vol. 8, Springer-Verlag, Berlin.

[164] Herman, G.T., Censor, Y., Gordon, D., and Lewitt, R. (1985) "Comment." (on the paper [258]), *Journal of the American Statistical Association* **80**, pp. 22–25.

[165] Herman, G. T. (1999) *private communication*.

[166] Herman, G. T. and Meyer, L. (1993) "Algebraic reconstruction techniques can be made computationally efficient." *IEEE Transactions on Medical Imaging* **12**, pp. 600–609.

[167] Hildreth, C. (1957) "A quadratic programming procedure." *Naval Research Logistics Quarterly* **4**, pp. 79–85. Erratum, p. 361.

[168] Hoffman, K., and Kunze, R. (1965) *Linear Algebra*. Prentice-Hall.

[169] Hogg, R. and Craig, A. (1978) *Introduction to Mathematical Statistics*, MacMillan, New York.

[170] Holte, S., Schmidlin, P., Linden, A., Rosenqvist, G. and Eriksson, L. (1990) "Iterative image reconstruction for positron emission tomography: a study of convergence and quantitation problems." *IEEE Transactions on Nuclear Science* **37**, pp. 629–635.

[171] Hudson, M., Hutton, B., and Larkin, R. (1992) "Accelerated EM reconstruction using ordered subsets." *Journal of Nuclear Medicine*, **33**, p.960.

[172] Hudson, H.M. and Larkin, R.S. (1994) "Accelerated image reconstruction using ordered subsets of projection data." *IEEE Transactions on Medical Imaging* **13**, pp. 601–609.

[173] Huesman, R., Klein, G., Moses, W., Qi, J., Ruetter, B., and Virador, P. (2000) "List-mode maximum likelihood reconstruction applied to positron emission mammography (PEM) with irregular sampling." *IEEE Transactions on Medical Imaging* **19 (5)**, pp. 532–537.

[174] Hutton, B., Kyme, A., Lau, Y., Skerrett, D., and Fulton, R. (2002) "A hybrid 3-D reconstruction/registration algorithm for correction of head motion in emission tomography." *IEEE Transactions on Nuclear Science* **49 (1)**, pp. 188–194.

[175] Jiang, M., and Wang, G. (2003) "Convergence studies on iterative algorithms for image reconstruction." *IEEE Transactions on Medical Imaging*, **22(5)**, pp. 569–579.

[176] Kaczmarz, S. (1937) "Angenäherte Auflösung von Systemen linearer Gleichungen." *Bulletin de l'Academie Polonaise des Sciences et Lettres* **A35**, pp. 355–357.

[177] Kak, A., and Slaney, M. (2001) *Principles of Computerized Tomographic Imaging*. SIAM, Philadelphia, PA.

[178] Kalman, R. (1960) "A new approach to linear filtering and prediction problems." *Trans. ASME, J. Basic Eng.* **82**, pp. 35–45.

[179] Katznelson, Y. (1983) *An Introduction to Harmonic Analysis*. New York: John Wiley and Sons, Inc.

[180] Kheifets, A. (2004) *private communication*.

[181] King, M., Glick, S., Pretorius, H., Wells, G., Gifford, H., Narayanan, M., and Farncombe, T. (2004) "Attenuation, scatter, and spatial resolution compensation in SPECT." in [263], pp. 473–498.

[182] Koltracht, L., and Lancaster, P. (1990) "Constraining strategies for linear iterative processes." *IMA J. Numer. Anal.*, **10**, pp. 555–567.

[183] Körner, T. (1988) *Fourier Analysis*. Cambridge, UK: Cambridge University Press.

[184] Körner, T. (1996) *The Pleasures of Counting*. Cambridge, UK: Cambridge University Press.

[185] Kuhn, H., and Tucker, A. (eds.) (1956) *Linear Inequalities and Related Systems*. Annals of Mathematical Studies, No. 38. New Jersey: Princeton University Press.

[186] Kullback, S. and Leibler, R. (1951) "On information and sufficiency." *Annals of Mathematical Statistics* **22**, pp. 79–86.

[187] Landweber, L. (1951) "An iterative formula for Fredholm integral equations of the first kind." *Amer. J. of Math.* **73**, pp. 615–624.

[188] Lane, R. (1987) "Recovery of complex images from Fourier magnitude." *Optics Communications* **63(1)**, pp. 6–10.

[189] Lange, K. and Carson, R. (1984) "EM reconstruction algorithms for emission and transmission tomography." *Journal of Computer Assisted Tomography* **8**, pp. 306–316.

[190] Lange, K., Bahn, M. and Little, R. (1987) "A theoretical study of some maximum likelihood algorithms for emission and transmission tomography." *IEEE Trans. Med. Imag.* **MI-6(2)**, pp. 106–114.

[191] La Rivière, P., and Vargas, P. (2006) "Monotonic penalized-likelihood image reconstruction for x-ray fluorescence computed tomography." *IEEE Transactions on Medical Imaging* **25(9)**, pp. 1117–1129.

[192] Leahy, R., Hebert, T., and Lee, R. (1989) "Applications of Markov random field models in medical imaging." in *Proceedings of the Conference on Information Processing in Medical Imaging* Lawrence-Berkeley Laboratory, Berkeley, CA.

[193] Leahy, R. and Byrne, C. (2000) "Guest editorial: Recent development in iterative image reconstruction for PET and SPECT." *IEEE Trans. Med. Imag.* **19**, pp. 257–260.

[194] Leis, A., Beck, M., Gruska, M., Best, C., Hegerl, R., Baumeister, W., and Leis, J. (2006) "Cryo-electron tomography of biological specimens." *IEEE Signal Processing Magazine*, **23 (3)**, pp. 95–103.

[195] Lent, A. (1998) *private communication*.

[196] Levitan, E. and Herman, G. (1987) "A maximum *a posteriori* probability expectation maximization algorithm for image reconstruction in emission tomography." *IEEE Transactions on Medical Imaging* **6**, pp. 185–192.

[197] Liao, C.-W., Fiddy, M., and Byrne, C. (1997) "Imaging from the zero locations of far-field intensity data." *Journal of the Optical Society of America -A* **14 (12)**, pp. 3155–3161.

[198] Luenberger, D. (1969) *Optimization by Vector Space Methods*. New York: John Wiley and Sons, Inc.

[199] Lustig, M., Donoho, D., and Pauly, J. (2008) *Magnetic Resonance in Medicine*, to appear.

[200] Magness, T., and McQuire, J. (1962) "Comparison of least squares and minimum variance estimates of regression parameters." *Annals of Mathematical Statistics* **33**, pp. 462–470.

[201] Mann, W. (1953) "Mean value methods in iteration." *Proc. Amer. Math. Soc.* **4**, pp. 506–510.

[202] McLachlan, G.J. and Krishnan, T. (1997) *The EM Algorithm and Extensions*. New York: John Wiley and Sons, Inc.

[203] McVeigh, E., and Ozturk, C. (2001) "Imaging myocardial strain." *IEEE Signal Processing Magazine*, **18 (6)**, pp. 44–56.

[204] Meidunas, E. (2001) "Re-scaled block iterative expectation maximization maximum likelihood (RBI-EMML) abundance estimation and sub-pixel material identification in hyperspectral imagery" *MS thesis, Department of Electrical Engineering, University of Massachusetts Lowell*.

[205] Meijering, E., Smal, I., and Danuser, G. (2006) "Tracking in molecular bioimaging." *IEEE Signal Processing Magazine*, **23 (3)**, pp. 46–53.

[206] Motzkin, T. and Schoenberg, I. (1954) "The relaxation method for linear inequalities." *Canadian Journal of Mathematics* **6**, pp. 393–404.

[207] Mumcuoglu, E., Leahy, R., and Cherry, S. (1996) "Bayesian reconstruction of PET images: Methodology and performance analysis." *Phys. Med. Biol.*, **41**, pp. 1777–1807.

[208] Narayanan, M., Byrne, C. and King, M. (2001) "An interior point iterative maximum-likelihood reconstruction algorithm incorporating upper and lower bounds with application to SPECT transmission imaging." *IEEE Transactions on Medical Imaging* **TMI-20 (4)**, pp. 342–353.

[209] Nash, S. and Sofer, A. (1996) *Linear and Nonlinear Programming*. New York: McGraw-Hill.

[210] Natterer, F. (1986) *Mathematics of Computed Tomography*. New York: John Wiley and Sons, Inc.

[211] Natterer, F., and Wübbeling, F. (2001) *Mathematical Methods in Image Reconstruction*. Philadelphia, PA: SIAM Publ.

[212] Ollinger, J., and Fessler, J. (1997) "Positron-emission tomography." *IEEE Signal Processing Magazine*, **14 (1)**, pp. 43–55.

[213] Oppenheim, A. and Schafer, R. (1975) *Digital Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall.

[214] Papoulis, A. (1975) "A new algorithm in spectral analysis and band-limited extrapolation." *IEEE Transactions on Circuits and Systems* **22**, pp. 735–742.

[215] Papoulis, A. (1977) *Signal Analysis*. New York: McGraw-Hill.

[216] Parra, L. and Barrett, H. (1998) "List-mode likelihood: EM algorithm and image quality estimation demonstrated on 2-D PET." *IEEE Transactions on Medical Imaging* **17**, pp. 228–235.

[217] Paulraj, A., Roy, R., and Kailath, T. (1986) "A subspace rotation approach to signal parameter estimation." *Proceedings of the IEEE* **74**, pp. 1044–1045.

[218] Peressini, A., Sullivan, F., and Uhl, J. (1988) *The Mathematics of Nonlinear Programming*. Berlin: Springer-Verlag.

[219] Peters, T. (1981) "Resolution improvement to CT systems using aperture-function correction." in [163], pp. 241–251.

[220] Pretorius, H., King, M., Pan, T-S, deVries, D., Glick, S., and Byrne, C. (1998) "Reducing the influence of the partial volume effect on SPECT activity quantitation with 3D modelling of spatial resolution in iterative reconstruction." *Phys.Med. Biol.* **43**, pp. 407–420.

[221] Pižurica, A., Philips, W., Lemahieu, I., and Acheroy, M. (2003) "A versatile wavelet domain noise filtration technique for medical imaging." *IEEE Transactions on Medical Imaging: Special Issue on Wavelets in Medical Imaging* **22**, pp. 323–331.

[222] Poggio, T. and Smale, S. (2003) "The mathematics of learning: dealing with data." *Notices of the American Mathematical Society* **50 (5)**, pp. 537–544.

[223] Priestley, M. B. (1981) *Spectral Analysis and Time Series*. Boston: Academic Press.

[224] Prony, G.R.B. (1795) "Essai expérimental et analytique sur les lois de la dilatabilité de fluides élastiques et sur celles de la force expansion de la vapeur de l'alcool, à différentes températures." *Journal de l'Ecole Polytechnique* (Paris) **1(2)**, pp. 24–76.

[225] Qi, J., Leahy, R., Cherry, S., Chatziioannou, A., and Farquhar, T. (1998) "High resolution 3D Bayesian image reconstruction using the microPET small animal scanner. " *Phys. Med. Biol.*, **43 (4)**, pp. 1001–1013.

[226] Qian, H. (1990) "Inverse Poisson transformation and shot noise filtering." *Rev. Sci. Instrum.* **61**, pp. 2088–2091.

[227] Quistgaard, J. (1997) "Signal acquisition and processing in medical diagnostic ultrasound." *IEEE Signal processing Magazine*, **14 (1)**, pp. 67–74.

[228] Reich, S. (1979) "Weak convergence theorems for nonexpansive mappings in Banach spaces." *Journal of Mathematical Analysis and Applications*, **67**, pp. 274–276.

[229] Rockafellar, R. (1970) *Convex Analysis*. Princeton, NJ: Princeton University Press.

[230] Rockmore, A., and Macovski, A. (1976) "A maximum likelihood approach to emission image reconstruction from projections." *IEEE Transactions on Nuclear Science*, **NS-23**, pp. 1428–1432.

[231] Sarder, P., and Nehorai, A. (2006) "Deconvolution methods for 3-D fluoresence microscopy images." *IEEE Signal Processing Magazine*, **23 (3)**, pp. 32–45.

[232] Saulnier, G., Blue, R., Newell, J., Isaacson, D., and Edic, P. (2001) "Electrical impedance tomography." *IEEE Signal Processing Magazine*, **18 (6)**, pp. 31–43.

[233] Schmidlin, P. (1972) "Iterative separation of sections in tomographic scintigrams." *Nucl. Med.* **15(1)**.

[234] Schmidt, R. (1981) "A signal subspace approach to multiple emitter location and spectral estimation." *PhD thesis, Stanford University*.

[235] Schultz, L., Blanpied, G., Borozdin, K., *et al.* (2007) "Statistical reconstruction for cosmic ray muon tomography." *IEEE Transactions on Image Processing*, **16(8)**, pp. 1985–1993.

[236] Shaw, C. (2010) "Dimensions in medical imaging: the more the better?" *Proceedings of the IEEE*, **98(1)**, pp. 2–5.

[237] Shepp, L., and Vardi, Y. (1982) "Maximum likelihood reconstruction for emission tomography." *IEEE Transactions on Medical Imaging*, **MI-1**, pp. 113–122.

[238] Shieh, M., Byrne, C., Testorf, M., and Fiddy, M. (2006) "Iterative image reconstruction using prior knowledge." *Journal of the Optical Society of America, A*, **23(6)**, pp. 1292–1300.

[239] Shieh, M., Byrne, C., and Fiddy, M. (2006) "Image reconstruction: a unifying model for resolution enhancement and data extrapolation: Tutorial." *Journal of the Optical Society of America, A*, **23(2)**, pp. 258–266.

[240] Shieh, M., and Byrne, C. (2006) "Image reconstruction from limited Fourier data." *Journal of the Optical Society of America, A*, **23(11)**.

[241] Smith, C. Ray and Grandy, W.T., editors (1985) *Maximum-Entropy and Bayesian Methods in Inverse Problems*. Dordrecht: Reidel Publ.

[242] Smith, C. Ray and Erickson, G., editors (1987) *Maximum-Entropy and Bayesian Spectral Analysis and Estimation Problems*. Dordrecht: Reidel Publ.

[243] Soares, E., Byrne, C., Glick, S., Appledorn, R., and King, M. (1993) "Implementation and evaluation of an analytic solution to the photon attenuation and nonstationary resolution reconstruction problem in SPECT." *IEEE Transactions on Nuclear Science*, **40 (4)**, pp. 1231–1237.

[244] Stark, H. and Yang, Y. (1998) *Vector Space Projections: A Numerical Approach to Signal and Image Processing, Neural Nets and Optics*. New York: John Wiley and Sons, Inc.

[245] Stiemke, E. (1915) "Über positive Lösungen homogener linearer Gleichungen." *Math. Ann*, **76**, pp. 340–342.

[246] Strang, G. (1980) *Linear Algebra and its Applications*. New York: Academic Press.

[247] Tanabe, K. (1971) "Projection method for solving a singular system of linear equations and its applications." *Numer. Math.* **17**, pp. 203–214.

[248] Therrien, C. (1992) *Discrete Random Signals and Statistical Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall.

[249] Thévenaz, P., Blu, T., and Unser, M. (2000) "Interpolation revisited." *IEEE Transactions on Medical Imaging*, **19**, pp.739–758.

[250] Tsui, B., Gullberg, G., Edgerton, E., Ballard, J., Perry, J., McCartney, W., and Berg, J. (1989) "Correction of non-uniform attenuation in cardiac SPECT imaging." *Journal of Nuclear Medicine*, **30(4)**, pp. 497–507.

[251] Tucker, A. (1956) "Dual systems of homogeneous linear relations." in [185], pp. 3–18.

[252] Twomey, S. (1996) *Introduction to the Mathematics of Inversion in Remote Sensing and Indirect Measurement*. New York: Dover Publ.

[253] Udpa, L., Ayres, V., Fan, Y., Chen, Q., Kumar, S. (2006) "Deconvolution of atomic force microscopy data for cellular and molecular imaging." *IEEE Signal Processing Magazine*, **23 (3)**, pp. 73–83.

[254] Unser, M. (1999) "Splines: A perfect fit for signal and image processing." *IEEE Signal Processing Magazine*, **16**, pp. 22–38.

[255] Van Trees, H. (1968) *Detection, Estimation and Modulation Theory*. New York: John Wiley and Sons, Inc.

[256] van der Sluis, A. (1969) "Condition numbers and equilibration of matrices." *Numer. Math.*, **14**, pp. 14–23.

[257] van der Sluis, A., and van der Vorst, H.A. (1990) "SIRT- and CG-type methods for the iterative solution of sparse linear least-squares problems." *Linear Algebra and its Applications*, **130**, pp. 257–302.

[258] Vardi, Y., Shepp, L.A. and Kaufman, L. (1985) "A statistical model for positron emission tomography." *Journal of the American Statistical Association* **80**, pp. 8–20.

[259] von Neumann, J., and Morgenstern, O. (1944) *Theory of Games and Economic Behavior*. New Jersey: Princeton University Press.

[260] von Neumann, J., and Goldstine, H. H. (1947) "Numerical inverting of matrices of high order." Bulletin of the American Mathematical Society, **53**, pp. 1021–1099.

[261] Vonesch, C., Aguet, F., Vonesch, J-L, and Unser, M. (2006) "The colored revolution in bio-imaging." *IEEE Signal Processing Magazine*, **23 (3)**, pp. 20–31.

[262] Weintraub, K. (2012) "Bloodless brain surgery." in *The Boston Globe*, April 23, 2012, page B5.

[263] Wernick, M. and Aarsvold, J., editors (2004) *Emission Tomography: The Fundamentals of PET and SPECT*. San Diego: Elsevier Academic Press.

[264] Wiener, N. (1949) *Time Series*. Cambridge, MA: MIT Press.

[265] Wright, G.A. (1997) "Magnetic resonance imaging." *IEEE Signal Processing Magazine*, **14 (1)**, pp. 56–66.

[266] Wright, M. (2009) "The dual flow between linear algebra and optimization." view-graphs of talk given at the History of Numerical Linear Algebra Minisymposium - Part II, SIAM Conference on Applied Linear Algebra, Monterey, CA, October 28, 2009.

[267] Wright, W., Pridham, R., and Kay, S. (1981) "Digital signal processing for sonar." *Proc. IEEE* **69**, pp. 1451–1506.

[268] Yang, Q. (2004) "The relaxed CQ algorithm solving the split feasibility problem." *Inverse Problems*, **20**, pp. 1261–1266.

[269] Yin, M. (2011) "About triangular matrices." , seminar notes.

[270] Yin, W., and Zhang, Y. (2008) "Extracting salient features from less data via l_1-minimization." *SIAG/OPT Views-and-News*, **19(1)**, pp. 11–19.

[271] Youla, D. (1978) "Generalized image restoration by the method of alternating projections." *IEEE Transactions on Circuits and Systems* **CAS-25 (9)**, pp. 694–702.

[272] Youla, D.C. (1987) "Mathematical theory of image restoration by the method of convex projections." in *Image Recovery: Theory and Applications*, pp. 29–78, Stark, H., editor (1987) Orlando FL: Academic Press.

[273] Young, R. (1980) *An Introduction to Nonharmonic Fourier Analysis*. Boston: Academic Press.

[274] Zhou, X., and Wong, S. (2006) "Informatics challenges of high-throughput microscopy." *IEEE Signal Processing Magazine*, **23 (3)**, pp. 63–72.

[275] Zimmer, C., Zhang, B., Dufour, A., Thébaud, A., Berlemont, S., Meas-Yedid, V., and Marin, J-C. (2006) "On the digital trail of mobile cells." *IEEE Signal Processing Magazine*, **23 (3)**, pp. 54–62.

Index