# Research Summaries

**Charles L. Byrne**
Department of Mathematical Sciences
University of Massachusetts Lowell
Lowell, MA 01854

March 22, 2012

(The most recent version is available as a pdf file at
http://faculty.uml.edu/cbyrne/cbyrne.html)

# Contents

# Chapter 1

# Introduction

## 1.1 Overview

In the course of doing my research I found that certain questions arise that, while probably not of much interest to anyone else, continue to nag me. I kept returning to these questions as a tongue explores a painful tooth. Recently, I had the good fortune to answer several of these questions, achieving what I think of as small private victories that will probably never be published. These questions often had to do with the relationship, if any, between different parts of my research. I decided that, since I will have more down time over the spring and summer of 2011 than I had expected to have, I would write up these private victories.

As I look back over the mathematics that I have done over the past several decades, I begin to see the outlines of a small number of themes, threads that seem to weave their way through what might seem largely unrelated work. This has prompted me to expand my effort and to try to capture some of those themes here.

The chapters that follow will trace aspects of my work in essentially chronological order, beginning about 1980. Some of the material has appeared in print previously. I will try to place the material in each chapter within the context of the problems and influences I was dealing with at the time.

Over the past couple of years I have also been putting together a collection of essays on various topics in applied mathematics, designed to supplement the text in my graduate courses. This effort involved research of a somewhat different sort, since I was not as familiar with some of these topics as I wanted to be. I have a few favorite ones that I am including here as well.

# Chapter 2

# The PDFT

## 2.1 The Context

In 1978 I received tenure at The Catholic University of America and began looking for new areas for research. Up to that time my research had been in functional analysis and topology, but without much focus. In the 1979-1980 academic year I taught a graduate course in Stochastic Processes out of Breiman; this was an area I knew only slightly and wished to know better. Ray Fitzgerald was in the class. He was a PhD physicist working in the Acoustics Division at the Naval Research Laboratory. We began by discussing some of the mathematics involved in the problems of interest to him and quickly moved on to a collaboration that lasted until about 1990. At first, ours was an informal arrangement, but after a short while I became a paid consultant, eventually spending the 1981-1982 and 1982-1983 academic years working at NRL, on leave-of-absence from CUA.

In 1982 Ray and I went to London and Paris for conferences and to meet with people from the UK Admiralty Research Laboratory in Teddington. At the meeting at Imperial College we met Mike Fiddy, then a physics professor at the University of London. We had already been in touch with Mike and his colleagues and had exchanges preprints. My meeting with Mike was the beginning of a collaboration that has continued to the present day.

Fourier analysis plays a central role in imaging farfield objects. As is discussed in detail in the book *A First Course in Signal Processing*, available on my website, what we can measure are often values of the Fourier transform of what we want. The amount of Fourier data available is usually limited and increasing the resolution is the main objective. Ray was interested in acoustic array signal processing, while Mike was involved in various applications of Fourier optics, but both wanted higher resolution. There had been some work by Gerchberg and Papoulis on finite, discretized

band-limited extrapolation, several other efforts in the direction of linear extrapolation, and the nonlinear, maximum entropy methods of Burg. Ray asked me to study these various methods and help him decide if they had any use in acoustic SP. The PDFT and the IPDFT were the results of this effort.

## 2.2 The Basic Problem

Suppose that $F(\omega)$ is in $L^2(-\pi, \pi)$ and its inverse Fourier transform is

$$f(x) = \frac{1}{2\pi} \int_{-\pi}^{\pi} F(\omega) e^{-ix\omega} d\omega. \tag{2.1}$$

In applications, $F(\omega)$ is unknown and we wish to reconstruct, or estimate, $F(\omega)$ from finitely many values of $f(x)$.

## 2.3 The DFT

To begin with, suppose that we have finitely many Nyquist samples, $f(n)$, for $n = -N, ..., N$. The *discrete Fourier transform* (DFT) estimate of $F(\omega)$ is

$$F_{DFT}(\omega) = \sum_{n=-N}^{N} f(n) e^{in\omega}. \tag{2.2}$$

This estimate is consistent with the data, in the sense that

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} F_{DFT}(\omega) e^{-in\omega} d\omega = f(n),$$

for $n = -N, ..., N$. Consequently, $F_{DFT}(\omega)$ might be the right answer. It is helpful to remember that the DFT estimate is the function of minimum $L^2$ norm that is consistent with the data.

A useful way to look at the problem is in the context of Hilbert space. The space $L^2(-\pi, \pi)$ has the inner product

$$\langle F, G \rangle = \frac{1}{2\pi} \int_{-\pi}^{\pi} F(\omega) \overline{G(\omega)} d\omega. \tag{2.3}$$

For each $n$ define

$$E_n(\omega) = e^{in\omega},$$

so that

$$f(n) = \langle F, E_n \rangle.$$

We know that $F(\omega)$ can be decomposed as

$$F(\omega) = F_{DFT}(\omega) + G(\omega), \tag{2.4}$$

where

$$\langle G, E_n \rangle = 0,$$

for $n = -N, ..., N$. So the process of taking the inner product of $F$ with the $E_n$ for $n = -N, ..., N$ can tell us nothing about the $G$. Many have concluded, wrongly, that the DFT is the best we can do, and that this means that we can never learn anything about $G$. The subtle point that they have missed is that we get to choose the ambient Hilbert space. Why must it be $L^2(-\pi, \pi)$?

## 2.4 The PDFT

Suppose that $\epsilon \leq P(\omega) \leq B$ is a positive function on $[-\pi, \pi]$ and we define a new inner product on $L^2(-\pi, \pi)$ by

$$\langle F, G \rangle_P = \frac{1}{2\pi} \int_{-\pi}^{\pi} F(\omega)\overline{G(\omega)}P(\omega)^{-1} d\omega. \tag{2.5}$$

For each function $G(\omega)$ in $L^2$ the square of the weighted norm is

$$\|G\|_P^2 = \int_{-\pi}^{\pi} |G(\omega)|^2 P(\omega)^{-1} d\omega = \langle G, G \rangle_P. \tag{2.6}$$

We can now write

$$f(n) = \langle F, H_n \rangle_P, \tag{2.7}$$

where

$$H_n(\omega) = P(\omega)E_n(\omega).$$

Every $F(\omega)$ can be decomposed as

$$F(\omega) = F_{PDFT}(\omega) + W(\omega), \tag{2.8}$$

where

$$F_{PDFT}(\omega) = \sum_{m=-N}^{N} a_m H_m(\omega) = P(\omega) \sum_{m=-N}^{N} a_m e^{im\omega}, \tag{2.9}$$

and

$$\langle W, H_n \rangle_P = 0,$$

for each $n = -N, ..., N$. Consequently,

$$\langle F, H_n \rangle_P = \langle F_{PDFT}, H_n \rangle_P,$$

for each $n = -N, ..., N$. We use this to find the coefficients $a_m$.

We have

$$f(n) = \frac{1}{2\pi} \sum_{m=-N}^{N} a_m \int_{-\pi}^{\pi} P(\omega) e^{i(m-n)\omega} d\omega, \tag{2.10}$$

for $n = -N, ..., N$. We solve this system of linear equations for the $a_m$ to get the PDFT estimate of $F(\omega)$. Unless $P(\omega)$ equals one throughout $[-\pi, \pi]$, the PDFT and the DFT are different estimators, although both are consistent with the data and either could be the right answer. The PDFT is the function consistent with the data for which the weighted norm is minimized.

The key point here, missed by many, is that we can represent the data using inner products in an infinite number of ways, each leading to a different data-consistent estimator of $F(\omega)$. The importance of the PDFT for applications lies in our ability to select $P(\omega)$ to incorporate features of the function $F(\omega)$ that may be known a priori.

## 2.5 Band-Limited Extrapolation

Suppose that $F(\omega) = 0$, except for $\omega$ in the interval $[-\Omega, \Omega]$, for some $\Omega < \pi$. We want an estimator of $F(\omega)$ that is consistent with the data and is also supported on the interval $[-\Omega, \Omega]$.

In 1979 I presented the MDFT [23], which turned out to be a special case of the PDFT, as a solution to the problem of band-limited extrapolation. The MDFT estimator has the form of a trig polynomial, restricted to the interval $[\Omega, \Omega]$; that is, the MDFT is

$$F_{MDFT}(\omega) = \chi_\Omega(\omega) \sum_{m=-N}^{N} a_m e^{im\omega}, \tag{2.11}$$

where $\chi_\Omega(\omega)$ is one for $|\omega| \leq \Omega$ and zero otherwise. We use the data consistency to find the coefficients. We have

$$f(n) = \frac{1}{2\pi} \sum_{m=-N}^{N} a_m \int_{-\Omega}^{\Omega} e^{i(m-n)\omega} d\omega, \tag{2.12}$$

for $n = -N, ..., N$. It turns out that this system can be ill-conditioned, particularly when $\Omega$ is much smaller than $\pi$. To reduce the sensitivity to noise, we always replace the function $\chi_\Omega(\omega)$ with

$$\chi_\Omega(\omega) + \epsilon \chi_\pi(\omega),$$

where $\epsilon > 0$ is a small positive quantity.

The inverse Fourier transform of $F_{MDFT}(\omega)$ is a function of $x$ that agrees with the data at the points $x = n$, $n = -N, ..., N$, and therefore provides a band-limited extrapolation of the data.

In the MDFT our prior knowledge that $F(\omega)$ is supported on the interval $[-\Omega, \Omega]$ is incorporated in the first factor, $\chi_\Omega(\omega)$. This suggests that, when we have other prior information about the overall shape of $F(\omega)$, we can incorporate that prior knowledge through the use of the function $P(\omega)$. In the next section we see how effective this can be.

## 2.6 Using More Prior Knowledge

An important point to keep in mind when doing signal processing 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 2.1 illustrates this point.

The images in Figure 2.1 were generated in 1983 at the University of London, by Angela Darling, one of Mike Fiddy's doctoral students. This one picture brought home to all of us how useful the PDFT could be.

The original image on the upper right of Figure 2.1 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-norm solution would seem to be a reasonable reconstruction method.

The minimum-norm solution (DFT) 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-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-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-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-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-norm reconstruction, using as the weight vector the reciprocals of the values of the prior image. This minimum-weighted-norm reconstruction is shown on the lower left; it is clearly almost the same as the original image.

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 affect 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.

## 2.7 Calculating the PDFT

We see from Equation (2.10) that we need to compute the values of the inverse Fourier transform of the function $P(\omega)$ at the points of the form $m-n$, for $m, n = -N, ..., N$. In practice, $N$ is often quite large and this step can be computationally expensive, particularly when $P(\omega)$ is not simple. Then we have to solve a large system of linear equations to obtain the coefficients. This whole process can be greatly simplified by discretizing the problem at the start. We replace the unknown function $F(\omega)$ and the functions $H_n(\omega)$ by finite vectors. The problem then is to find the minimum weighted norm solution of a system of linear equations. All we need is a discrete version of the function $P(\omega)$, not its inverse Fourier transform values. The solution can be obtained using the iterative ART algorithm [141].

## 2.8 Using the PDFT

As we have seen, the PDFT is the unique function, consistent with the data, for which the weighted two-norm is minimized. So long as the support of $P(\omega)$ is larger than that of the true $F(\omega)$, the weighted two-norm of the PDFT will not be excessively large, since the weighted two-norm of $F(\omega)$ itself is (typically) not large. But if the support of $P(\omega)$ is smaller than that of the true $F(\omega)$ there need not be any reasonable function with this smaller support that is also data-consistent. In such cases, the weighted two-norm of the PDFT will typically be quite large. We can often use this fact to estimate the true support of $F(\omega)$.

In [28, 29] we applied similar reasoning to solve the phase problem, in which we have not values $f(n)$, but $|f(n)|$. Our idea here was to select phases iteratively to go with the magnitude data, checking each time to see how large the weighted two-norm of the resulting PDFT was. When the phases were wrong, there need not be any reasonable function consistent with this constructed data and having the given support. But when the phases were close enough to the correct ones, the weighted two-norm of the PDFT began to drop. We found that it was not necessary to get the missing phases exactly; reasonably good choices for the phases sufficed to generate good images.

## 2.9 The PDFT and Minimum One-Norm Solutions

The PDFT applies more generally to under-determined systems of linear equations $Ax = b$. In such cases we can find the minimum two-norm solution, the minimum weighted two-norm solution (e.g. the PDFT), and the minimum one-norm solution.

The minimum one-norm solution is the $x$ for which

$$||x||_1 = \sum_{n=1}^{N} |x_n|$$

is minimized, subject to $Ax = b$. Denote the solution by $x^*$. This problem can be formulated as a linear programming problem, so is more easily solved.

### 2.9.1 Minimum One-Norm as an LP Problem

The entries of $x$ need not be non-negative, so the problem is not yet a linear programming problem. Let

$$B = [\, A \quad -A \,],$$

and consider the linear programming problem of minimizing the function

$$c^T z = \sum_{j=1}^{2J} z_j,$$

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 $u_j^* v_j^* = 0$, for each $j$. If, say, there is a $j$ such that $0 < v_j^* < u_j^*$, then we can create a new vector $z$ by replacing the old $u_j^*$ with $u_j^* - v_j^*$ and the old $v_j^*$ with zero, while maintaining $Bz = b$. But then, since $u_j^* - v_j^* < u_j^* + v_j^*$, 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 $u_j = x_j$, if $x_j \geq 0$, and $u_j = 0$, otherwise. Let $v_j = u_j - x_j$, so that $x = u - v$. Then let

$$z = \begin{bmatrix} u \\ v \end{bmatrix}.$$

Then $b = Ax = Bz$, and $c^T z = \|x\|_1$. Consequently,

$$\|x^*\|_1 = c^T z^* \leq c^T z = \|x\|_1,$$

and $x^*$ must be a minimum one-norm solution.

### 2.9.2 Why the One-Norm?

When a system of linear equations $Ax = b$ is under-determined, we can find the *minimum-two-norm solution* that minimizes the square of the two-norm,

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

subject to $Ax = b$. 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 solution for which the one-norm,

$$||x||_1 = \sum_{n=1}^{N} |x_n|,$$

is minimized. The one-norm still penalizes relatively large entries $x_n$ more than the smaller ones, but much less than the two-norm does.

### 2.9.3 Comparison with the PDFT

The generalized PDFT approach to solving the under-determined system $Ax = b$ is to select weights $w_n > 0$ and then to find the solution $\tilde{x}$ that minimizes the weighted two-norm given by

$$\sum_{n=1}^{N} |x_n|^2 w_n.$$

Our intention is to select weights $w_n$ so that $w_n^{-1}$ is reasonably close to $|x_n^*|$; 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, $\tilde{x}$ also minimizes the one-norm.

### 2.9.4 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 PDFT solution using these weights, then using this PDFT 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 [86].

In [55], 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 PDFT concerns the nature of the prior weighting. Does $w_n$ approximate $|x_n|^{-1}$ or $|x_n|^{-2}$? This is close to the issue treated in [55], 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 PDFT, 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 $\hat{x}$, their use reduces the penalty associated with the larger entries of an estimated solution.

## 2.10 Summary

From a purely mathematical standpoint, the PDFT is not particularly deep. As I see it, the PDFT contributed in three ways to the overall problem of reconstruction: first, it has turned out to be quite a useful technique; second, it has provided the basis for other reconstruction methods, such as in the phase problem [28], as well as the nonlinear IPDFT, to be discussed in the next chapter; and third, it embodied a somewhat novel philosophy of reconstruction, by which we are freed from the dominance of the $L^2$ formulation and allowed to formulate the problem in any convenient Hilbert space.

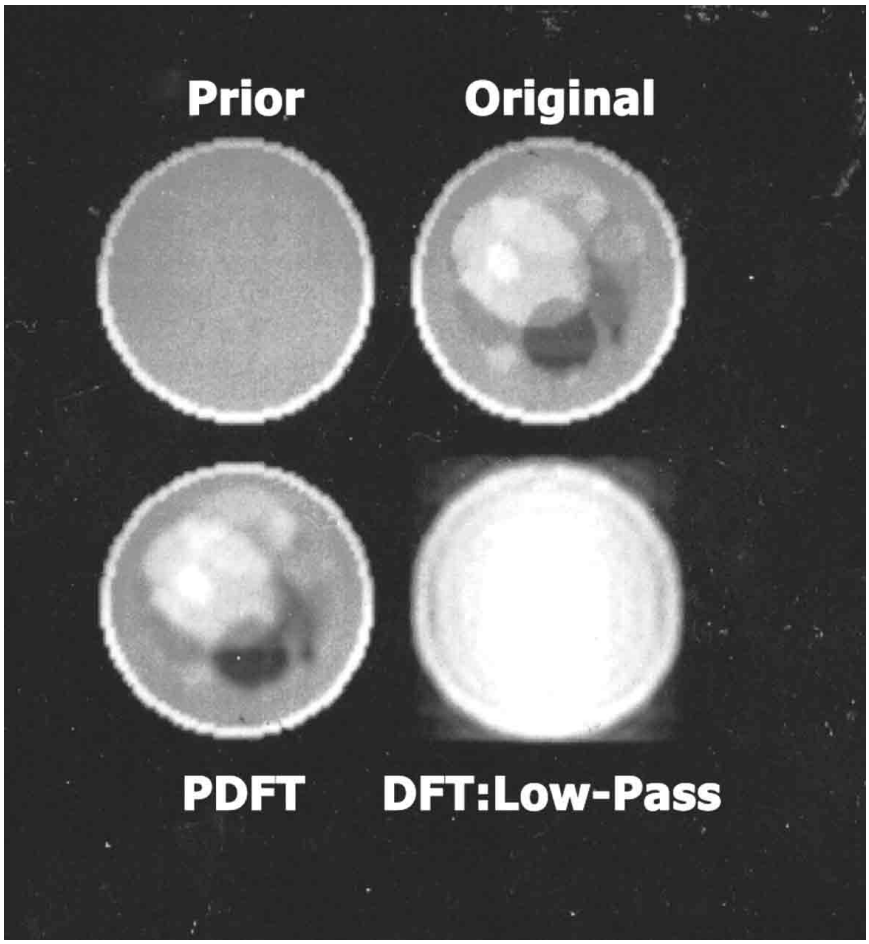


Figure 2.1: Extracting information in image reconstruction.

# Chapter 3

# The IPDFT

## 3.1 The Context

In a variety of applications, the unknown function $F(\omega)$ is non-negative and consists of a small number of delta functions, embedded in a (possibly flat) background. Linear reconstruction methods may not be able to resolve all of the delta functions. To solve this problem nonlinear, high resolution methods can be used. Throughout this chapter we use $R(\omega)$ and $r(x)$ instead of $F(\omega)$ and $f(x)$ as a reminder that the unknown function we seek is non-negative now. Since $R(\omega)$ is non-negative, we know that $r(-x) = \overline{r(x)}$, so it is natural that we suppose that our data is $r(n)$ for $|n| \leq N$.

In the 1970's, John Burg, working in the oil industry, developed his *maximum entropy* method (MEM). This spawned a tremendous interest in entropy and nonlinear reconstruction methods, and led to the two conferences in Laramie in the early 1980's [145, 146]. When I began consulting for NRL, Ray Fitzgerald asked me to look into these high-resolution methods, to see what use they may have for SONAR. I attended the first of the two Laramie conferences. The IPDFT, which is a generalization of the MEM, appeared in [26].

## 3.2 Burg's MEM

The problem of estimating the nonnegative function $R(\omega)$, for $|\omega| \leq \pi$, from the finitely many Fourier-transform values

$$r(n) = \int_{-\pi}^{\pi} R(\omega) \exp(-in\omega) d\omega/2\pi, \; n = -N, ..., N$$

is an *under-determined problem*, meaning that the data alone is insufficient to determine a unique answer. In such situations we must select one so-

lution out of the infinitely many that are mathematically possible. The obvious questions we need to answer are: What criteria do we use in this selection? How do we find algorithms that meet our chosen criteria? In this chapter we consider Burg's *maximum entropy method* (MEM) [18, 19].

These values $r(n)$ are autocorrelation function values associated with a random process having $R(\omega)$ for its power spectrum. In many applications, such as seismic remote sensing, these autocorrelation values are estimates obtained from relatively few samples of the underlying random process, so that $N$ is not large. The DFT estimate,

$$R_{DFT}(\omega) = \sum_{n=-N}^{N} r(n) \exp(in\omega),$$

is real-valued and consistent with the data, but is not necessarily nonnegative. For small values of $N$, the DFT may not be sufficiently resolving to be useful. This suggests that one criterion we can use to perform our selection process is to require that the method provide better resolution than the DFT for relatively small values of $N$, when reconstructing power spectra that consist mainly of delta functions.

The objective of Burg's MEM for estimating a power spectrum is to seek better resolution by combining nonnegativity and data-consistency in a single closed-form estimate. The MEM is remarkable in that it is the only closed-form (that is, non-iterative) estimation method that is guaranteed to produce an estimate that is both non-negative and consistent with the autocorrelation samples. Later we shall consider a more general method, the inverse PDFT (IPDFT), that is both data-consistent and positive in most cases.

In discussing the Burg MEM estimate, we shall need to refer to the concept of *minimum-phase* vectors. We consider that briefly now.

We say that the finite column vector with complex entries $(a_0, a_1, ..., a_N)^T$ is a *minimum-phase* vector if the complex polynomial

$$A(z) = a_0 + a_1 z + ... + a_N z^N$$

has the property that $A(z) = 0$ implies that $|z| > 1$; that is, all roots of $A(z)$ are outside the unit circle. Consequently, the function $B(z)$ given by $B(z) = 1/A(z)$ is analytic in a disk centered at the origin and including the unit circle. Therefore, we can write

$$B(z) = b_0 + b_1 z + b_2 z^2 + ...,$$

and taking $z = \exp(i\omega)$, we get

$$B(\exp(i\omega)) = b_0 + b_1 \exp(i\omega) + b_2 \exp(2i\omega) + ... \ .$$

The point here is that $B(\exp(i\omega))$ is a one-sided trigonometric series, with only terms corresponding to $\exp(in\omega)$ for nonnegative $n$.

The MEM approach is to estimate $R(\omega)$ by the function $S(\omega) > 0$ that maximizes the so-called Burg entropy, $\int_{-\pi}^{\pi} \log S(\theta) d\theta$, subject to the data constraints.

The Euler-Lagrange equation from the calculus of variations allows us to conclude that $S(\omega)$ has the form

$$S(\omega) = 1/H(\omega)$$

for

$$H(\omega) = \sum_{n=-N}^{N} h_n e^{in\omega} > 0.$$

From the Fejér-Riesz Theorem we know that $H(\omega) = |A(e^{i\omega})|^2$ for minimum phase $A(z)$. As we now show, the coefficients $a_n$ satisfy a system of linear equations formed using the data $r(n)$.

Given the data $r(n), |n| \leq N$, we form the *autocorrelation matrix* $R$ with entries $R_{mn} = r(m-n)$, for $-N \leq m, n \leq N$. Let $\delta$ be the column vector $\delta = (1, 0, ..., 0)^T$. Let $\mathbf{a} = (a_0, a_1, ..., a_N)^T$ be the solution of the system $R\mathbf{a} = \delta$. Then, Burg's MEM estimate is the function $S(\omega) = R_{MEM}(\omega)$ given by

$$R_{MEM}(\omega) = a_0/|A(\exp(i\omega))|^2, |\omega| \leq \pi.$$

Once we show that $a_0 \geq 0$, it will be obvious that $R_{MEM}(\omega) \geq 0$. We also must show that $R_{MEM}$ is data-consistent; that is,

$$r(n) = \int_{-\pi}^{\pi} R_{MEM}(\omega) \exp(-in\omega) d\omega/2\pi =, \; n = -N, ..., N.$$

Let us write $R_{MEM}(\omega)$ as a Fourier series; that is,

$$R_{MEM}(\omega) = \sum_{n=-\infty}^{+\infty} q(n) \exp(in\omega), \; |\omega| \leq \pi.$$

From the form of $R_{MEM}(\omega)$, we have

$$R_{MEM}(\omega)\overline{A(\exp(i\omega))} = a_0 B(\exp(i\omega)). \tag{3.1}$$

It can be shown that $A(z)$ has all its roots outside the unit circle, so $B(\exp(i\omega))$ is a one-sided trigonometric series, with only terms corresponding to $\exp(in\omega)$ for nonnegative $n$. Then, multiplying on the left side of Equation (3.1), and equating coefficients corresponding to $n = 0, -1, -2, ...$, we find that, provided $q(n) = r(n)$, for $|n| \leq N$, we must have $R\mathbf{a} = \delta$. Notice that these are precisely the same equations we solve in calculating

the coefficients of an AR process. For that reason the MEM is sometimes called an autoregressive method for spectral estimation.

The MEM resolves better than the DFT when the true power spectrum being reconstructed is a sum of delta functions plus a flat background. When the background itself is not flat, performance of the MEM degrades rapidly; the MEM tends to interpret any non-flat background in terms of additional delta functions. In the next section we consider an extension of the MEM, called the indirect PDFT (IPDFT), that corrects this flaw.

## 3.3 The IPDFT

The IPDFT method is suggested by considering the MEM system of equations $R\mathbf{a} = \delta$ and comparing it with the linear system that arises in the PDFT. In the PDFT the matrix of the system comes from our prior estimate $P(\omega)$, the right side of the equation is the data vector, and we solve for the vector of coefficients. The PDFT estimate then has two factors, the prior $P(\omega)$ and the finite trig polynomial. When we view the MEM system this way, it appears that $R(\omega)$ is playing the role of the prior. The data then consists of Fourier coefficients of a constant function.

If we try to estimate this constant function using $R(\omega)$ as our prior, we would expect the finite trig polynomial factor to correspond essentially to the reciprocal of $R(\omega)$. Said another way, we would expect $R(\omega)$ to be well approximated by the reciprocal of the finite trig polynomial. When $R(\omega)$ consists of a flat background plus a few delta functions, the trig polynomial should remove from $R(\omega)$ everything that is not flat, namely the delta functions. It does so by placing its zeros very near the supports of the delta functions.

Suppose now that $R(\omega)$ consists of delta functions on top of a non-flat background. The trig polynomial will now remove from $R(\omega)$ everything that is not flat, which means that the trig polynomial will approximate the reciprocal of the non-flat background. The zeros of the trig polynomial will be placed near the high-intensity areas of the background. The MEM estimate will then have a number of spikes in the regions of high-intensity of the background, making it difficult to find the true delta functions.

In the IPDFT we replace the vector $\delta$ on the right side of the system $R\mathbf{a} = \delta$ with the vector

$$\mathbf{p} = (p(0), p(1), ..., p(N))^T,$$

where $p(x)$ is the inverse Fourier transform of the prior $P(\omega)$. Now the trig polynomial removes everything from $R(\omega)$ that does not look like $P(\omega)$, which should be only the delta functions, if $P(\omega)$ is a good approximation of the background. For more details and a comparison of MEM and IPDFT see [44].

## 3.4 Technical Issues

In our discussion of the MEM, we obtained an estimate for the function $R(\omega)$, not simply a way of locating the delta-function components. As we shall show, the IPDFT can also be used to estimate $R(\omega)$. Although the resulting estimate is not guaranteed to be nonnegative or data consistent, it usually is both of these.

The equations that we solve in the IPDFT are

$$p(m) = \sum_{k=0}^{N} f_k r(m-k). \tag{3.2}$$

Once we have found $\mathbf{f}$ we form the polynomial

$$F(\omega) = \sum_{k=0}^{N} f_k e^{ik\omega},\ |\omega| \leq \pi.$$

The zeros of $F(\omega)$ should then be near the supports of the delta function components of the power spectrum $R(\omega)$, provided that our original estimate of the background is not too inaccurate.

For any function $G(\omega)$ on $[-\pi, \pi]$ with Fourier series

$$G(\omega) = \sum_{n=-\infty}^{\infty} g(n) e^{in\omega},$$

the *additive causal part* of the function $G(\omega)$ is

$$G_+(\omega) = \sum_{n=0}^{\infty} g(n) e^{in\omega}.$$

Any function such as $G_+$ that has Fourier coefficients that are zero for negative indices is called a *causal function*. The Equation (3.2) then says that the two causal functions $P_+$ and $(FR)_+$ have Fourier coefficients that agree for $m = 0, 1, ..., N$.

Because $F(\omega)$ is a finite causal trigonometric polynomial, we can write

$$(FR)_+(\omega) = R_+(\omega)F(\omega) + J(\omega),$$

where

$$J(\omega) = \sum_{m=0}^{N-1} \Big( \sum_{k=1}^{N-m} r(-k) f_{m+k} \Big) e^{im\omega}.$$

Treating $P_+$ as approximately equal to $(FR)_+ = R_+F + J$, we obtain as an estimate of $R_+$ the function $Q = (P_+ - J)/F$. In order for this estimate

of $R_+$ to be causal, it is sufficient that the function $1/F$ be causal. This means that the trigonometric polynomial $F(\omega)$ must be minimum phase; that is, all its roots lie outside the unit circle. We know that this is always the case for MEM. It is not always the case for the IPDFT, but it is usually the case in practice; in fact, it was difficult (but possible) to construct a counterexample. We then construct our IPDFT estimate of $R(\omega)$, which is

$$R_{IPDFT}(\omega) = 2\text{Re}(Q(\omega)) - r(0).$$

The IPDFT estimate is real-valued and, when $1/F$ is causal, guaranteed to be data consistent. Although this estimate is not guaranteed to be nonnegative, it usually is.

We know that the vector $\mathbf{a}$ that solves $R\mathbf{a} = \delta$ corresponds to a polynomial $A(z)$ having all its roots on or outside the unit circle; that is, it is minimum phase. The IPDFT involves the solution of the system $R\mathbf{f} = \mathbf{p}$, where $\mathbf{p} = (p(0), ..., p(N))^T$ is the vector of initial Fourier coefficients of another power spectrum, $P(\omega) \geq 0$ on $[-\pi, \pi]$. When $P(\omega)$ is constant, we get $\mathbf{p} = \delta$. For the IPDFT to be data-consistent, it is sufficient that the polynomial $F(z) = f_0 + ... + f_N z^N$ be minimum phase. Although this need not be the case, it is usually observed in practice.

## 3.5 Afterward

My collaboration with Mike Fiddy and his graduate students in London, which began in 1982, led to several trips to London, to his spending a sabbatical year with me in 1985, and then with his moving to UML in 1987. Our work in the 1980's involved various applications of the PDFT, especially the phase problem.

In 1983 Alan Steele came to NRL from Adelaide and we began a collaboration that continued through my visit to Australia in June 1986. The work focused mainly on stabilizing eigenvector techniques.

In about 1983 NRL opened a second branch in Bay Saint Louis, MS. In 1987 Don DelBalzo, who had moved from DC to MS, asked Applied Technologies, Inc. to hire me to give a three-day short course on SONAR signal processing. This was the first of several short courses I gave on this subject over the next five years in the US and Canada. For several years after the MS short course I collaborated with DelBalzo and his colleagues on matched-field processing.

My research was to take a different path after I was invited to visit the UMass Medical School in 1989.

# Chapter 4

# A Tale of Two Algorithms

## 4.1 The Context

In 1989 I began consulting for the Nuclear Medicine group at the University of Massachusetts Medical School. At that time the leaders of the group, Drs. Mike King and Bill Penney, were working on SPECT image reconstruction and asked me to study what I will call here the EMML algorithm. This algorithm is a particular case of the more general EM algorithm [74]. Some of the material in this chapter first appeared in [30, 31].

In [30] I established the close connection between the EMML and the SMART algorithms, answered a couple of open questions, and corrected some mistakes that had appeared in the literature. After the publication of [30] I was invited by Rob Lewitt to speak at the Medical Imaging Processing Group (MIPG) at Penn, where I met Gabor Herman, the head of MIPG, Yair Censor, Arnold Lent and Paul Eggermont. A bit later, Yair invited me to Haifa, to participate in a conference at the Technion.

I was also invited by Larry Shepp to participate in a week-long IMA conference at the University of Minnesota. The presentation in this chapter follows closely my talk at that conference, which was published in [32]. It is taken from one of my texts, which explains the embedded exercises. Actually, I purposely presented many of the results in the form of exercises to emphasize the elementary nature of this proof.

## 4.2 Background

In positron emission tomography (PET) and single-photon emission computed tomography (SPECT) radioactive material, or radionuclide, is injected into the patient and is then metabolized. Photons exiting the body are detected at various locations and this data provides the basis for a

reconstruction of the distribution of the radionuclide. The radionuclide is designed to provide a contrast between, say, a tumor, and healthy nearby tissue. In the fully discrete model, the body consists of pixels or voxels, each with an unknown amount of the radionuclide. For simplicity, the average number of emissions coming from a given pixel during the scanning time is taken as a surrogate for the actual amount of radionuclide present in a pixel. The goal is to determine these average numbers.

In 1976 Rockmore and Macovski [137] suggested that Poisson statistics be used and the average number of emissions be viewed as parameters to be determined, for example, by maximum likelihood estimation. Shepp and Vardi [139, 149] took the next step and presented the EMML algorithm for this particular case. A complete and elementary proof of convergence of the EMML first appeared in [30].

At the same time, Gabor Herman, Yair Censor, and their colleagues were performing image reconstruction using an algebraic approach, involving the solution of large, constrained systems of linear equations. They had introduced the ART and MART algorithms, and suggested, in [99], their comments on the paper [149], that the EMML was probably closely related to their algebraic methods. This prompted a heated response from the authors of [149], who denied any connection between their statistical method and any linear-algebraic approach. The *simultaneous* MART (SMART) [71, 138] is a variant of MART that uses all the data at each step of the iteration. It was my development of the EMML and SMART in tandem [30, 31, 32] and my presentation at the IMA that revealed just how close the two methods really are.

Although the EMML and SMART algorithms have quite different histories and are not typically considered together, they are closely related, as we shall see [30, 31]. In this chapter we examine these two algorithms in tandem, following [32]. 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 [30] was based on the analogous proof for the EMML [149], while discovery of the faster version of the EMML, the *rescaled block-iterative* EMML (RBI-EMML) [33] came from studying the analogous block-iterative version of SMART [59]. The proofs we give here are elementary and rely mainly on easily established properties of the cross-entropy or Kullback-Leibler distance.

Another class of iterative algorithms was introduced into medical imaging by Gordon et al. in [93]. 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.

## 4.3 The Kullback-Leibler Distance

The Kullback-Leibler distance $KL(a,b)$ is defined for positive $a$ and $b$ by

$$KL(a,b) = a\log\frac{a}{b} + b - a, \tag{4.1}$$

with $KL(a,0) = +\infty$ and $KL(0,b) = b$. The KL distance is then extended to non-negative vectors $\mathbf{x}$ and $\mathbf{z}$ component-wise;

$$KL(\mathbf{x},\mathbf{z}) = \sum_{n=1}^{N} KL(x_n, z_n). \tag{4.2}$$

Clearly, the KL distance has the property

$$KL(c\mathbf{x}, c\mathbf{z}) = cKL(\mathbf{x},\mathbf{z})$$

for all positive scalars $c$.

**Ex. 4.1** *Let* $z_+ = \sum_{j=1}^{J} z_j > 0$. *Then*

$$KL(\mathbf{x},\mathbf{z}) = KL(x_+, z_+) + KL(\mathbf{x}, (x_+/z_+)\mathbf{z}). \tag{4.3}$$

As we shall see, the KL distance mimics the ordinary Euclidean distance in several ways that make it particularly useful in designing optimization algorithms.

## 4.4 The Alternating Minimization Paradigm

Let $P$ be an $I$ by $J$ matrix with entries $P_{ij} \geq 0$, such that, for each $j = 1, ..., J$, we have $s_j = \sum_{i=1}^{I} P_{ij} > 0$. Let $\mathbf{y} = (y_1, ..., y_I)^T$ with $y_i > 0$ 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 $P_{ij}$ with $P_{ij}/s_j$; the quantities $(P\mathbf{x})_i$ are unchanged.

For each nonnegative vector $\mathbf{x}$ for which $(P\mathbf{x})_i = \sum_{j=1}^{J} P_{ij}x_j > 0$, let $r(\mathbf{x}) = \{r(\mathbf{x})_{ij}\}$ and $q(\mathbf{x}) = \{q(\mathbf{x})_{ij}\}$ be the $I$ by $J$ arrays with entries

$$r(\mathbf{x})_{ij} = x_j P_{ij}\frac{y_i}{(P\mathbf{x})_i}$$

and

$$q(\mathbf{x})_{ij} = x_j P_{ij}.$$

The KL distances

$$KL(r(\mathbf{x}), q(\mathbf{z})) = \sum_{i=1}^{I} \sum_{j=i}^{J} KL(r(\mathbf{x})_{ij}, q(\mathbf{z})_{ij})$$

and

$$KL(q(\mathbf{x}), r(\mathbf{z})) = \sum_{i=1}^{I} \sum_{j=1}^{J} KL(q(\mathbf{x})_{ij}, r(\mathbf{z})_{ij})$$

will play important roles in the discussion that follows. Note that if there is nonnegative $\mathbf{x}$ with $r(\mathbf{x}) = q(\mathbf{x})$ then $\mathbf{y} = P\mathbf{x}$.

### 4.4.1 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 $KL(r(\mathbf{x}), q(\mathbf{z}))$ and $KL(q(\mathbf{x}), r(\mathbf{z}))$ are minimized, first with respect to the variable $\mathbf{x}$ and then with respect to the variable $\mathbf{z}$. 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. 4.2** *Establish the following Pythagorean identities:*

$$KL(r(\mathbf{x}), q(\mathbf{z})) = KL(r(\mathbf{z}), q(\mathbf{z})) + KL(r(\mathbf{x}), r(\mathbf{z})); \tag{4.4}$$

$$KL(r(\mathbf{x}), q(\mathbf{z})) = KL(r(\mathbf{x}), q(\mathbf{x}')) + KL(\mathbf{x}', \mathbf{z}), \tag{4.5}$$

*for*

$$x'_j = x_j \sum_{i=1}^{I} P_{ij} \frac{y_i}{(P\mathbf{x})_i}; \tag{4.6}$$

$$KL(q(\mathbf{x}), r(\mathbf{z})) = KL(q(\mathbf{x}), r(\mathbf{x})) + KL(\mathbf{x}, \mathbf{z}) - KL(P\mathbf{x}, P\mathbf{z}); \tag{4.7}$$

$$KL(q(\mathbf{x}), r(\mathbf{z})) = KL(q(\mathbf{z}''), r(\mathbf{z})) + KL(\mathbf{x}, \mathbf{z}''), \tag{4.8}$$

*for*

$$z''_j = z_j \exp(\sum_{i=1}^{I} P_{ij} \log \frac{y_i}{(P\mathbf{z})_i}). \tag{4.9}$$

*Note that it follows from Equation (4.3) that* $KL(\mathbf{x}, \mathbf{z}) - KL(P\mathbf{x}, P\mathbf{z}) \geq 0$.

### 4.4.2 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 $\mathbf{y} = P\mathbf{x}$ has nonnegative solutions, both algorithms produce such a solution. In general, the EMML gives a nonnegative minimizer of $KL(\mathbf{y}, P\mathbf{x})$, while the SMART minimizes $KL(P\mathbf{x}, \mathbf{y})$ over nonnegative $\mathbf{x}$.

For both algorithms we begin with an arbitrary positive vector $\mathbf{x}^0$. The iterative step for the EMML method is

$$x_j^{k+1} = (\mathbf{x}^k)'_j = x_j^k \sum_{i=1}^{I} P_{ij} \frac{y_i}{(P\mathbf{x}^k)_i}. \tag{4.10}$$

The iterative step for the SMART is

$$x_j^{m+1} = (\mathbf{x}^m)''_j = x_j^m \exp\Big(\sum_{i=1}^{I} P_{ij} \log \frac{y_i}{(P\mathbf{x}^m)_i}\Big). \tag{4.11}$$

Note that, to avoid confusion, we use $k$ for the iteration number of the EMML and $m$ for the SMART.

**Ex. 4.3** *Show that, for $\{\mathbf{x}^k\}$ given by Equation (4.10), $\{KL(\mathbf{y}, P\mathbf{x}^k)\}$ is decreasing and $\{KL(\mathbf{x}^{k+1}, \mathbf{x}^k)\} \to 0$. Show that, for $\{\mathbf{x}^m\}$ given by Equation (4.11), $\{KL(P\mathbf{x}^m, \mathbf{y})\}$ is decreasing and $\{KL(\mathbf{x}^m, \mathbf{x}^{m+1})\} \to 0$.*

**Hint:** Use $KL(r(\mathbf{x}), q(\mathbf{x})) = KL(\mathbf{y}, P\mathbf{x})$, $KL(q(\mathbf{x}), r(\mathbf{x})) = KL(P\mathbf{x}, \mathbf{y})$, and the Pythagorean identities.

**Ex. 4.4** *Show that the EMML sequence $\{\mathbf{x}^k\}$ is bounded by showing*

$$\sum_{j=1}^{J} x_j^k = \sum_{i=1}^{I} y_i.$$

*Show that the SMART sequence $\{\mathbf{x}^m\}$ is bounded by showing that*

$$\sum_{j=1}^{J} x_j^m \leq \sum_{i=1}^{I} y_i.$$

**Ex. 4.5** *Show that* $(\mathbf{x}^*)' = \mathbf{x}^*$ *for any cluster point* $\mathbf{x}^*$ *of the EMML sequence* $\{\mathbf{x}^k\}$ *and that* $(\mathbf{x}^*)'' = \mathbf{x}^*$ *for any cluster point* $\mathbf{x}^*$ *of the SMART sequence* $\{\mathbf{x}^m\}$.

**Hint:** Use the facts that $\{KL(\mathbf{x}^{k+1}, \mathbf{x}^k)\} \to 0$ and $\{KL(\mathbf{x}^m, \mathbf{x}^{m+1})\} \to 0$.

**Ex. 4.6** *Let* $\hat{\mathbf{x}}$ *and* $\tilde{\mathbf{x}}$ *minimize* $KL(\mathbf{y}, P\mathbf{x})$ *and* $KL(P\mathbf{x}, \mathbf{y})$, *respectively, over all* $\mathbf{x} \geq \mathbf{0}$. *Then,* $(\hat{\mathbf{x}})' = \hat{\mathbf{x}}$ *and* $(\tilde{\mathbf{x}})'' = \tilde{\mathbf{x}}$.

**Hint:** Apply Pythagorean identities to $KL(r(\hat{\mathbf{x}}), q(\hat{\mathbf{x}}))$ and $KL(q(\tilde{\mathbf{x}}), r(\tilde{\mathbf{x}}))$.

Note that, because of convexity properties of the KL distance, even if the minimizers $\hat{\mathbf{x}}$ and $\tilde{\mathbf{x}}$ are not unique, the vectors $P\hat{\mathbf{x}}$ and $P\tilde{\mathbf{x}}$ are unique.

**Ex. 4.7** *For the EMML sequence* $\{\mathbf{x}^k\}$ *with cluster point* $\mathbf{x}^*$ *and* $\hat{\mathbf{x}}$ *as defined previously, we have the double inequality*

$$KL(\hat{\mathbf{x}}, \mathbf{x}^k) \geq KL(r(\hat{\mathbf{x}}), r(\mathbf{x}^k)) \geq KL(\hat{\mathbf{x}}, \mathbf{x}^{k+1}), \tag{4.12}$$

*from which we conclude that the sequence* $\{KL(\hat{\mathbf{x}}, \mathbf{x}^k)\}$ *is decreasing and* $KL(\hat{\mathbf{x}}, \mathbf{x}^*) < +\infty$.

**Hint:** For the first inequality calculate $KL(r(\hat{\mathbf{x}}), q(\mathbf{x}^k))$ in two ways. For the second one, use $(\mathbf{x})'_j = \sum_{i=1}^I r(\mathbf{x})_{ij}$ and Exercise 4.1.

**Ex. 4.8** *Show that, for the SMART sequence* $\{\mathbf{x}^m\}$ *with cluster point* $\mathbf{x}^*$ *and* $\tilde{\mathbf{x}}$ *as defined previously, we have*

$$KL(\tilde{\mathbf{x}}, \mathbf{x}^m) - KL(\tilde{\mathbf{x}}, \mathbf{x}^{m+1}) = KL(P\mathbf{x}^{m+1}, \mathbf{y}) - KL(P\tilde{\mathbf{x}}, \mathbf{y}) +$$

$$KL(P\tilde{\mathbf{x}}, P\mathbf{x}^m) + KL(\mathbf{x}^{m+1}, \mathbf{x}^m) - KL(P\mathbf{x}^{m+1}, P\mathbf{x}^m), \tag{4.13}$$

*and so* $KL(P\tilde{\mathbf{x}}, P\mathbf{x}^*) = 0$, *the sequence* $\{KL(\tilde{\mathbf{x}}, \mathbf{x}^m)\}$ *is decreasing and* $KL(\tilde{\mathbf{x}}, \mathbf{x}^*) < +\infty$.

**Hint:** Expand $KL(q(\tilde{\mathbf{x}}), r(\mathbf{x}^m))$ using the Pythagorean identities.

**Ex. 4.9** *For $\mathbf{x}^*$ a cluster point of the EMML sequence $\{\mathbf{x}^k\}$ we have $KL(\mathbf{y}, P\mathbf{x}^*) = KL(\mathbf{y}, P\hat{\mathbf{x}})$. Therefore, $\mathbf{x}^*$ is a nonnegative minimizer of $KL(\mathbf{y}, P\mathbf{x})$. Consequently, the sequence $\{KL(\mathbf{x}^*, \mathbf{x}^k)\}$ converges to zero, and so $\{\mathbf{x}^k\} \to \mathbf{x}^*$.*

**Hint:** Use the double inequality of Equation (4.12) and $KL(r(\hat{\mathbf{x}}), q(\mathbf{x}^*))$.

**Ex. 4.10** *For $\mathbf{x}^*$ a cluster point of the SMART sequence $\{\mathbf{x}^m\}$ we have $KL(P\mathbf{x}^*, \mathbf{y}) = KL(P\tilde{\mathbf{x}}, \mathbf{y})$. Therefore, $\mathbf{x}^*$ is a nonnegative minimizer of $KL(P\mathbf{x}, \mathbf{y})$. Consequently, the sequence $\{KL(\mathbf{x}^*, \mathbf{x}^m)\}$ converges to zero, and so $\{\mathbf{x}^m\} \to \mathbf{x}^*$. Moreover,*

$$KL(\tilde{\mathbf{x}}, \mathbf{x}^0) \geq KL(\mathbf{x}^*, \mathbf{x}^0)$$

*for all $\tilde{\mathbf{x}}$ as before.*

**Hints:** Use Exercise 4.8. For the final assertion use the fact that the difference $KL(\tilde{\mathbf{x}}, \mathbf{x}^m) - KL(\tilde{\mathbf{x}}, \mathbf{x}^{m+1})$ is independent of the choice of $\tilde{\mathbf{x}}$, since it depends only on $P\mathbf{x}^* = P\tilde{\mathbf{x}}$. 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.

## 4.5 Related Topics

The idea of alternating minimization (altmin) that we use here is studied in great detail in the paper by Csiszár and Tusnády [70]. As the authors of [149] remark, the geometric argument in [70] is "deep, though it is hard to follow". One of my private victories recently has been to come to a better understanding of this paper and to obtain a somewhat simpler treatment of the altmin method, which I shall discuss in a later chapter.

As I noted earlier, the EMML algorithm is a particular case of the more general EM algorithm discussed in [74]. I have been studying the EM algorithm for about twenty years and have been bothered all that time by the erroneous manner in which most articles and books treat the case of continuous-variable pdf's.

In 2009 Paul Eggermont and I were invited to write a chapter on the EM algorithm for a book on algorithms. This assignment brought me back yet again to the problem that has bothered me all these years. Earlier this year, I discovered what I believe to be the answer to this problem, which I will also discuss in a later chapter.

# Chapter 5

# The Rescaled Block-Iterative Method

## 5.1 The Context

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; in tomography, there can be tens of thousands of equations. In the early 1990's Hudson, Hutton and Larkin [103, 104] discovered, partly by accident, that useful images can be reconstructed when only some of the equations are used at each step of the algorithm. They called this approach the *ordered-subset* (OSEM) variation of the EMML algorithm. The OSEM was picked up quickly by the research community and soon became the main concern of those studying medical image reconstruction algorithms.

Ordered-subset methods, also known as 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. These methods should always converge to a single solution, when the data is noise-free; the OSEM, however, does not always do that. The proof of convergence of the OSEM holds only when the subsets exhibit rather special properties, called *subset balance*. I suspected that OSEM is not the final word on block-iterative extensions of the EMML.

In the fall of 1995 I had a sabbatical and spent much of that time searching for a correct version of the OSEM. Eventually, I discovered what I called the *rescaled block-iterative* EMML (RBI-EMML) [33]. As it should, the RBI-EMML converges for noise-free data, for any choice of subsets. When subset balance holds, the RBI-EMML reduces to OSEM.

Even though the OSEM is not the mathematically correct algorithm, it works well enough so that the medical imaging community, already invested in OSEM, saw little reason to start using RBI-EMML. Consequently, I am forced to consider the discovery of the RBI-EMML one of my more or less private victories.

## 5.2 Recalling the MART Algorithm

Throughout this chapter $A$ will denote a rectangular matrix with non-negative entries, $b$ a vector with positive entries, and $x$ an unknown vector with non-negative entries. The MART algorithm uses only one equation at a time. For $k = 0, 1, ...$, we let $i = k(\mathrm{mod}\, I) + 1$ and

$$x_j^{k+1} = x_j^k \Big(\frac{b_i}{(Ax^k)_i}\Big)^{A_{ij}m_i^{-1}}. \tag{5.1}$$

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

## 5.3 The EMML and the SMART Algorithms

We recall the formulas for the iterative step of the EMML and the SMART.

### 5.3.1 The EMML Algorithm

The iterative step for the EMML algorithm is

$$x_j^{k+1} = x_j^k s_j^{-1} \sum_{i=1}^{I} A_{ij} \frac{b_i}{(Ax^k)_i}, \tag{5.2}$$

where $s_j = \sum_{i=1}^{I} A_{ij}$. The iterative step can also be written as

$$x_j^{k+1} = \sum_{i=1}^{I} (s_j^{-1} A_{ij}) \Big(x_j^k \frac{b_i}{(Ax^k)_i}\Big), \tag{5.3}$$

which shows that $x_j^{k+1}$ is the weighted arithmetic mean of the terms $x_j^k \frac{b_i}{(Ax^k)_i}$.

### 5.3.2 The SMART Algorithm

The iterative step for the SMART algorithm is

$$x_j^{k+1} = x_j^k \exp\Big(s_j^{-1} \sum_{i=1}^{I} A_{ij} \log\Big(\frac{b_i}{(Ax^k)_i}\Big)\Big). \tag{5.4}$$

The iterative step can also be written as

$$x_j^{k+1} = \prod_{i=1}^{I} \Big( x_j^k \frac{b_i}{(Ax^k)_i} \Big)^{s_j^{-1} A_{ij}}, \tag{5.5}$$

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.

## 5.4 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 $n$th 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 $i$th equation to compute $x^{k+1}$.

### 5.4.1 Block-Iterative SMART

More general block-iterative versions of the SMART algorithm have been known since the work of Darroch and Ratcliff [71], and were treated in detail in [59]. The iterative step of the block-iterative SMART (BI-SMART) algorithm is

$$x_j^{k+1} = x_j^k \exp\Big( m_n^{-1} \sum_{i \in B_n}^{I} A_{ij} \log\Big( \frac{b_i}{(Ax^k)_i} \Big) \Big). \tag{5.6}$$

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) [33].

### 5.4.2 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.* [102]. Somewhat later, Hudson, Hutton and Larkin [103, 104] discovered what they called the *ordered-subset* (OSEM) variation of the EMML algorithm.

The iterative step of the OSEM algorithm is

$$x_j^{k+1} = x_j^k s_{nj}^{-1} \sum_{i \in B_n} A_{ij} \Big( \frac{b_i}{(Ax^k)_i} \Big). \tag{5.7}$$

It is identical with that of the EMML in Equation (5.2), 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.

### 5.4.3 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 (5.1) as

$$x_j^{k+1} = \Big( x_j^k \Big)^{1 - A_{ij} m_i^{-1}} \Big( x_j^k \frac{b_i}{(Ax^k)_i} \Big)^{A_{ij} m_i^{-1}}. \tag{5.8}$$

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

$$x_j^{k+1} = (x_j^k)^{1-m_n^{-1}s_{nj}} \prod_{i \in B_n} \Big( x_j^k \frac{b_i}{(Ax^k)_i} \Big)^{A_{ij}m_n^{-1}}; \tag{5.9}$$

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

$$x_j^{k+1} = (1 - m_n^{-1}s_{nj})x_j^k + m_n^{-1}x_j^k \sum_{i \in B_n} A_{ij} \Big( \frac{b_i}{(Ax^k)_i} \Big). \tag{5.10}$$

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) [33]. 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.

### 5.4.4 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

$$x_j^{k+1} = (1 - m_i^{-1}A_{ij})x_j^k + m_i^{-1}A_{ij} \Big( x_j^k \frac{b_i}{(Ax^k)_i} \Big). \tag{5.11}$$

## 5.5 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 $H_i$.

## 5.6 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 6

# The Split Feasibility Problem

## 6.1 The Context

In the late 1990's I began working more closely with Yair Censor, studying iterative algorithms more from a mathematical point of view, rather than for applications [38, 40, 21]. The split feasibility problem (SFP) was something Yair and Tommy Elfving had been working on for about a decade. Their algorithms for the SFP, as well as one I came up with, were impractical. Finally, I was able to use the approach in [38] to come up with what I called the CQ algorithm [41]. Because it did not require a nested inversion at each step, it was practical.

The CQ algorithm can be viewed as an iterative method for optimizing a convex function. As I pursued this idea further, I found that I was able to extend it to some variational inequality problems. I emailed several people to see if such extensions were known. Boris Polyak replied that Dolidze had done something similar, and this led me to a new way of looking at the CQ algorithm, as I explained in [42].

Shortly after the appearance of [41] and [42] Yair, in collaboration with some medical physicists, extended the CQ algorithm and applied it to intensity modulated radiation therapy [58, 57]. At the same time, a number of other researchers, seemingly all from China, began publishing various extensions of the CQ algorithm. As a result, the paper [42] became quite popular and got referenced frequently. In 2010 the journal *Inverse Problems* celebrated their 25th anniversary with a special issue, containing one paper from each of their first twenty-five years. The paper [42] was chosen to represent 2004; a not-so-private victory this time.

## 6.2 The Split Feasibility Problem

The *split feasibility problem* (SFP) [56] 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 $R^J$ and $R^I$, respectively. When there is no exact solution to the SFP the CQ algorithm optimizes a certain proximity measure. In this chapter we discuss the CQ algorithm for solving the SFP, as well as recent extensions and applications.

## 6.3 The CQ Algorithm

In [41] the CQ algorithm for solving the SFP was presented, for the real case. It has the iterative step

$$x^{k+1} = P_C(x^k - \gamma A^T(I - P_Q)Ax^k), \tag{6.1}$$

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 $C^J$ and $C^I$, respectively. The iterative step of the extended CQ algorithm is then

$$x^{k+1} = P_C(x^k - \gamma A^\dagger(I - P_Q)Ax^k). \tag{6.2}$$

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

$$f(x) = \frac{1}{2}||P_QAx - Ax||_2^2 \tag{6.3}$$

over the set $C$, provided such constrained minimizers exist. Therefore the CQ algorithm is an iterative constrained optimization method. As shown in [42], convergence of the CQ algorithm is a consequence of the Krasnoselskii-Mann (KM) Theorem for averaged operators (see [45, 50]).

The function $f(x)$ is convex and differentiable on $R^J$ and its derivative is the operator

$$\nabla f(x) = A^T(I - P_Q)Ax; \tag{6.4}$$

see [3].

The following lemma contains terms not defined here; the interested reader should consult [50].

**Lemma 6.1** *The derivative operator $\nabla f$ is $\lambda$-Lipschitz continuous for $\lambda = \rho(A^TA)$, therefore it is $\nu$-ism for $\nu = \frac{1}{\lambda}$.*

**Proof:** We have

$$||\nabla f(x) - \nabla f(y)||_2^2 = ||A^T(I - P_Q)Ax - A^T(I - P_Q)Ay||_2^2 \tag{6.5}$$

$$\leq \lambda||(I - P_Q)Ax - (I - P_Q)Ay||_2^2. \tag{6.6}$$

Also

$$||(I - P_Q)Ax - (I - P_Q)Ay||_2^2 = ||Ax - Ay||_2^2 \tag{6.7}$$

$$+||P_QAx - P_QAy||_2^2 - 2\langle P_QAx - P_QAy, Ax - Ay\rangle \tag{6.8}$$

and, since $P_Q$ is fne,

$$\langle P_QAx - P_QAy, Ax - Ay\rangle \geq ||P_QAx - P_QAy||_2^2. \tag{6.9}$$

Therefore,

$$||\nabla f(x) - \nabla f(y)||_2^2 \leq \lambda(||Ax - Ay||_2^2 - ||P_QAx - P_QAy||_2^2) \tag{6.10}$$

$$\leq \lambda||Ax - Ay||_2^2 \leq \lambda^2||x - y||_2^2. \tag{6.11}$$

This completes the proof. ∎

If $\gamma \in (0, 2/\lambda)$ then $B = P_C(I - \gamma A^T(I - P_Q)A)$ is av and, by the KM Theorem, the orbit sequence $\{B^kx\}$ converges to a fixed point of $B$, whenever such points exist. If $z$ is a fixed point of $B$, then we have

$$z = P_C(z - \gamma A^T(I - P_Q)Az).$$

Therefore, for any $c$ in $C$ we have

$$\langle c - z, z - (z - \gamma A^T(I - P_Q)Az)\rangle \geq 0. \tag{6.12}$$

This tells us that

$$\langle c - z, A^T(I - P_Q)Az\rangle \geq 0, \tag{6.13}$$

which means that $z$ minimizes $f(x)$ relative to the set $C$.

The CQ algorithm employs the relaxation parameter $\gamma$ in the interval $(0, 2/L)$, where $L$ is the largest eigenvalue of the matrix $A^TA$. Choosing the best relaxation parameter in any algorithm is a nontrivial procedure. Generally speaking, we want to select $\gamma$ near to $1/L$. We have a simple estimate for $L$ in the case of singular values of sparse matrices: if $A$ is normalized so that each row has length one, then the spectral radius of $A^TA$ does not exceed the maximum number of nonzero elements in any column of $A$. A similar upper bound on $\rho(A^TA)$ was obtained for non-normalized, $\epsilon$-sparse $A$.

## 6.4 Particular Cases of the CQ Algorithm

It is easy to find important examples of the SFP: if $C \subseteq R^J$ and $Q = \{b\}$ then solving the SFP amounts to solving the linear system of equations $Ax = b$; if $C$ is a proper subset of $R^J$, 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 [112] and projected Landweber methods (see [12]), are particular cases of the CQ algorithm.

### 6.4.1 The Landweber algorithm

With $x^0$ arbitrary and $k = 0, 1, ...$ let

$$x^{k+1} = x^k + \gamma A^T(b - Ax^k). \tag{6.1}$$

This is the Landweber algorithm.

### 6.4.2 The Projected Landweber Algorithm

For a general nonempty closed convex $C$, $x^0$ arbitrary, and $k = 0, 1, ...$, the projected Landweber method for finding a solution of $Ax = b$ in $C$ has the iterative step

$$x^{k+1} = P_C(x^k + \gamma A^T(b - Ax^k)). \tag{6.2}$$

### 6.4.3 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.

Another example of the CQ algorithm is the *simultaneous algebraic reconstruction technique* (SART) of Anderson and Kak for solving $Ax = b$, for nonnegative matrix $A$ [2].

### 6.4.4 Related Methods and Applications

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 [155].

In a recent papers [58, 57] Censor *et al.* discuss the application of the CQ algorithm to the problem of intensity-modulated radiation therapy (IMRT) treatment planning. Mathematically speaking, the problem is the *multi-set split feasibility problem* (MSSFP), which is to find $x$ in $C$, the non-empty intersection of closed, convex sets $C_i$, for $i = 1, ..., I$, such that $Ax$ is in the non-empty intersection of the closed, convex sets $Q_j$, for $j = 1, ..., J$. In the CQ algorithm it is assumed that the orthogonal projections onto $C$ and $Q$ are easily calculated, while algorithms for solving the MSSFP assume that the orthogonal projections onto the $C_i$ and $Q_j$ are easily calculated.

The split feasibility problem can be formulated as an optimization problem, namely, to minimize

$$h(x) = \psi_C(x) + \psi_Q(Ax), \tag{6.3}$$

where $\psi_C(x)$ is the indicator function of the set $C$. The CQ algorithm solves the more general problem of minimizing the function

$$f(x) = \psi_C(x) + ||P_Q Ax - Ax||_2^2. \tag{6.4}$$

The second term in $f(x)$ is differentiable, allowing us to apply the forward-backward splitting method of Combettes and Wajs [67]; the CQ algorithm is then a special case of their method.

## 6.5 Exercises

**Ex. 6.1** *Use the CQ algorithm to prove the following. Let $C_1$ and $C_2$ be nonempty, closed convex sets in $R^J$, with $C_1 \cap C_2 = \emptyset$. Assume that there is a unique $\hat{c}_2$ in $C_2$ minimizing the function $f(x) = ||c_2 - P_1 c_2||_2$, over all $c_2$ in $C_2$. Let $\hat{c}_1 = P_1 \hat{c}_2$. Then $P_2 \hat{c}_1 = \hat{c}_2$. Let $z^0$ be arbitrary and, for $n = 0, 1, ...$, let*

$$z^{2n+1} = P_1 z^{2n}, \tag{6.5}$$

*and*

$$z^{2n+2} = P_2 z^{2n+1}. \tag{6.6}$$

*Then*

$$\{z^{2n+1}\} \to \hat{c}_1, \tag{6.7}$$

*and*

$$\{z^{2n}\} \to \hat{c}_2. \tag{6.8}$$

# Chapter 7

# Sequential Unconstrained Minimization- SUMMA

## 7.1 The Context

In this chapter we consider an approach to optimization in which the original problem is replaced by a series of simpler problems. This approach can be particularly effective for constrained optimization. Suppose, for example, that we want to minimize $f(x)$, subject to the constraint that $x$ lie within a set $C$. At the $k$th step of the iteration we minimize the function $G_k(x) = f(x) + g_k(x)$, with no additional restrictions on $x$, to get the vector $x^k$, where the functions $g_k(x)$ are related to the set $C$ in some way. In practice, minimizing $G_k(x)$ may require iteration, but we will not deal with that issue here. In the best case, the sequence $\{x^k\}$ will converge to the solution to the original problem.

Several of the algorithms I have worked on fall into this category. In the case of the EMML algorithm, the function we want to minimize is $f(x) = KL(y, Px)$, but at each step we minimize something else, say $G_k(x) = f(x) + g_k(x)$, where each $G_k(x)$ has a minimizer that we can write in closed form. While the function $f(x)$ does not, by its very form, require that $x$ be non-negative, the function $G_k(x)$ does. So we achieve two things by using this approach: first, we get a closed form solution $x^k$ at each step; and second, the $x^k$ is automatically positive, which forces the limit to be non-negative. Much the same can be said of the SMART.

The EMML and SMART algorithms can be sensitive to noise in the data. For that reason, one usually adds a penalty term to regularize the problem. This is yet another reason for using sequential unconstrained minimization.

In [38] I considered the problem of sequential projection onto convex

sets using different metrics for each set. This situation arises in the EMML and SMART and I wondered if such an algorithm could be possible in general. It turned out that you can do it, provided that each projection is relaxed in a particular way. This again leads to a sequential unconstrained minimization procedure.

In 2007, with all these different examples of sequential unconstrained minimization in mind, I began looking at what sort of conditions we would need to place on the $g_k(x)$ to guarantee convergence. Out of this came what I called the SUMMA, which is a very general class of sequential unconstrained methods that, remarkably, includes just about everything. The SUMMA condition is so simple,

$$0 \leq g_{k+1}(x) \leq G_k(x) - G_k(x^k),$$

and yet quite powerful. It is odd, though, that try as I might, I cannot remember just how I happened to hit on this condition.

Among other things, the SUMMA contained the SMART as a particular case. However, as of the writing of [46], I could not see how to place the EMML under the SUMMA umbrella. It wasn't until 2011, when I saw how to include altmin as SUMMA, that the answer for EMML became clear.

## 7.2 Introduction

In many inverse problems, we have measured data pertaining to the object $x$, which may be, for example, a vectorized image, as well as prior information about $x$, such as that its entries are nonnegative. Tomographic imaging is a good example. We want to find an estimate of $x$ that is (more or less) consistent with the data, as well as conforming to the prior constraints. The measured data and prior information are usually not sufficient to determine a unique $x$ and some form of optimization is performed. For example, we may seek the image $x$ for which the entropy is maximized, or a minimum-norm least-squares solution.

There are many well-known methods for minimizing a function $f : R^J \rightarrow R$; we can use the Newton-Raphson algorithm or any of its several approximations, or nonlinear conjugate-gradient algorithms, such as the Fletcher-Reeves, Polak-Ribiere, or Hestenes-Stiefel methods. When the problem is to minimize the function $f(x)$, subject to constraints on the variable $x$, the problem becomes much more difficult. For such constrained minimization, we can employ *sequential unconstrained minimization algorithms* [85].

We assume that $f : S \rightarrow (-\infty, +\infty]$; our objective is to minimize $f(x)$ over $x$ in some given nonempty subset $P$ of $S$. At the $k$th step of a sequential unconstrained minimization algorithm we minimize a function $G_k(x)$ to get the vector $x^k$. We shall assume throughout that $G_k(x)$ has a

global minimizer $x^k$, for each $k$. The existence of these minimizers can be established, once additional conditions, such as continuity and convexity, are placed on the functions $G_k(x)$; see, for example, Fiacco and McCormick [85], p.95. Later we shall consider briefly the issue of computing the $x^k$.

In the best case, the set $S$ is a metric space and the sequence $\{x^k\}$ converges to a constrained minimizer of the original objective function $f(x)$. Obviously, the functions $G_k(x)$ must involve both the function $f(x)$ and the set $P$. Those methods for which each $x^k$ is *feasible*, that is, each $x^k$ is in $P$, are called *interior-point* methods, while those for which only the limit of the sequence is in $P$ are called *exterior-point* methods. Barrier-function algorithms are typically interior-point methods, while penalty-function algorithms are exterior-point methods. The purpose of this chapter is to present a fairly broad class of sequential unconstrained minimization algorithms, which we call SUMMA [46]. The SUMMA include both barrier-function algorithms, as well as proximity-function methods of Teboulle [148] and Censor and Zenios [60, 61], and the simultaneous multiplicative algebraic reconstruction technique (SMART) and the EMML algorithm [30, 43, 44, 45], and all alternating minimization methods for which the three- and four-point properties hold [70]. After some reformulation, the penalty-function methods can also be viewed as belonging to the SUMMA class.

## 7.3 SUMMA

The sequential unconstrained minimization algorithms (SUMMA) we present here use functions of the form

$$G_k(x) = f(x) + g_k(x), \tag{7.1}$$

with the auxiliary functions $g_k(x)$ chosen so that

$$0 \leq g_{k+1}(x) \leq G_k(x) - G_k(x^k), \tag{7.2}$$

for $k = 1, 2, ....$ Let

$$d = \inf\{f(x)|x \in P\} \geq -\infty,$$

and $x^k \in P$ for each $k$. Our main results are that the sequence $\{f(x^k)\}$ is monotonically decreasing to $d$, and, subject to certain conditions on $S$ and the function $f(x)$, the sequence $\{x^k\}$ converges to a feasible $x^*$ with $f(x^*) = d$.

We begin with a brief review of several types of sequential unconstrained minimization methods, including those mentioned previously. Then we state and prove the convergence results for the SUMMA. Finally, we show that each of these methods reviewed previously is a particular case of the SUMMA.

## 7.4 Barrier-Function Methods (I)

Let $b(x) : R^J \rightarrow (-\infty, +\infty]$ be continuous, with effective domain the set

$$D = \{x |\, b(x) < +\infty\}.$$

The goal is to minimize the objective function $f(x)$, over $x$ in the closed set $C = \overline{D}$, the closure of $D$. In the barrier-function method, we minimize

$$f(x) + \frac{1}{k} b(x) \tag{7.3}$$

over $x$ in $D$ to get $x^k$. Each $x^k$ lies within $D$, so the method is an interior-point algorithm. If the sequence $\{x^k\}$ converges, the limit vector $x^*$ will be in $C$ and $f(x^*) = f(\hat{x})$.

Barrier functions typically have the property that $b(x) \rightarrow +\infty$ as $x$ approaches the boundary of $D$, so not only is $x^k$ prevented from leaving $D$, it is discouraged from approaching the boundary.

### 7.4.1 Examples of Barrier Functions

Consider the convex programming (CP) problem of minimizing the convex function $f : R^J \rightarrow R$, subject to $g_i(x) \leq 0$, where each $g_i : R^J \rightarrow R$ is convex, for $i = 1, ..., I$. Let $D = \{x | g_i(x) < 0, i = 1, ..., I\}$; then $D$ is open. We consider two barrier functions appropriate for this problem.

#### The Logarithmic Barrier Function

A suitable barrier function is the *logarithmic barrier function*

$$b(x) = \Big( -\sum_{i=1}^{I} \log(-g_i(x)) \Big). \tag{7.4}$$

The function $-\log(-g_i(x))$ is defined only for those $x$ in $D$, and is positive for $g_i(x) > -1$. If $g_i(x)$ is near zero, then so is $-g_i(x)$ and $b(x)$ will be large.

#### The Inverse Barrier Function

Another suitable barrier function is the *inverse barrier function*

$$b(x) = \sum_{i=1}^{I} \frac{-1}{g_i(x)}, \tag{7.5}$$

defined for those $x$ in $D$.

In both examples, when $k$ is small, the minimization pays more attention to $b(x)$, and less to $f(x)$, forcing the $g_i(x)$ to be large negative numbers. But, as $k$ grows larger, more attention is paid to minimizing $f(x)$ and the $g_i(x)$ are allowed to be smaller negative numbers. By letting $k \to \infty$, we obtain an iterative method for solving the constrained minimization problem.

**An Illustration**

We minimize the function $f(x_1, x_2) = x_1^2 + x_2^2$, subject to the constraint that $x_1 + x_2 \geq 1$. The constraint is then written $g(x_1, x_2) = 1 - (x_1 + x_2) \leq 0$. We use the logarithmic barrier. The vector $x^k = (x_1^k, x_2^k)$ minimizing the function

$$G_k(x) = x_1^2 + x_2^2 - \frac{1}{k}\log(x_1 + x_2 - 1)$$

has entries

$$x_1^k = x_2^k = \frac{1}{4} + \frac{1}{4}\sqrt{1 + \frac{4}{k}}.$$

Notice that $x_1^k + x_2^k > 1$, so each $x^k$ satisfies the constraint. As $k \to +\infty$, $x^k$ converges to $(\frac{1}{2}, \frac{1}{2})$, which is the solution to the original problem. An obvious question is why we have used a minus sign rather than a plus sign in front of the logarithm. If we had used a plus sign there would have been no minimizer of $G_k(x)$ within the feasible region of all $(x_1, x_2)$ with $x_1 + x_2 \geq 1$.

## 7.5 Penalty-Function Methods (I)

Instead of minimizing a function $f(x)$ over $x$ in $R^J$, we sometimes want to minimize a *penalized* version, $f(x) + p(x)$. As with barrier-function methods, the new function $f(x) + p(x)$ may be the function we really want to minimize, and we still need to find a method for doing this. In other cases, it is $f(x)$ that we wish to minimize, and the inclusion of the term $p(x)$ occurs only in the iterative steps of the algorithm. As we shall see, under conditions to be specified later, the penalty-function method can be used to minimize a continuous function $f(x)$ over the nonempty set of minimizers of another continuous function $p(x)$.

### 7.5.1 Imposing Constraints

When we add a barrier function to $f(x)$ we restrict the domain. When the barrier function is used in a sequential unconstrained minimization algorithm, the vector $x^k$ that minimizes the function $f(x) + \frac{1}{k}b(x)$ lies in the effective domain $D$ of $b(x)$, and we prove that, under certain conditions,

the sequence $\{x^k\}$ converges to a minimizer of the function $f(x)$ over the closure of $D$. The constraint of lying within the set $\overline{D}$ is satisfied at every step of the algorithm; for that reason such algorithms are called interior-point methods. Constraints may also be imposed using a penalty function. In this case, violations of the constraints are discouraged, but not forbidden. When a penalty function is used in a sequential unconstrained minimization algorithm, the $x^k$ need not satisfy the constraints; only the limit vector need be feasible.

### 7.5.2 Examples of Penalty Functions

Consider the CP problem. We wish to minimize the convex function $f(x)$ over all $x$ for which the convex functions $g_i(x) \leq 0$, for $i = 1, ..., I$.

#### The Absolute-Value Penalty Function

We let $g_i^+(x) = \max\{g_i(x), 0\}$, and

$$p(x) = \sum_{i=1}^{I} g_i^+(x). \tag{7.6}$$

This is the *Absolute-Value* penalty function; it penalizes violations of the constraints $g_i(x) \leq 0$, but does not forbid such violations. Then, for $k = 1, 2, ...$, we minimize

$$f(x) + kp(x), \tag{7.7}$$

to get $x^k$. As $k \to +\infty$, the penalty function becomes more heavily weighted, so that, in the limit, the constraints $g_i(x) \leq 0$ should hold. Because only the limit vector satisfies the constraints, and the $x^k$ are allowed to violate them, such a method is called an *exterior-point* method.

#### The Courant-Beltrami Penalty Function

The *Courant-Beltrami* penalty-function method is similar, but uses

$$p(x) = \sum_{i=1}^{I} [g_i^+(x)]^2. \tag{7.8}$$

#### The Quadratic-Loss Penalty Function

Penalty methods can also be used with equality constraints. Consider the problem of minimizing the convex function $f(x)$, subject to the constraints $g_i(x) = 0$, $i = 1, ..., I$. The *quadratic-loss* penalty function is

$$p(x) = \frac{1}{2} \sum_{i=1}^{I} (g_i(x))^2. \tag{7.9}$$

The inclusion of a penalty term can serve purposes other than to impose constraints on the location of the limit vector. In image processing, it is often desirable to obtain a reconstructed image that is locally smooth, but with well defined edges. Penalty functions that favor such images can then be used in the iterative reconstruction [87]. We survey several instances in which we would want to use a penalized objective function.

**Regularized Least-Squares**

Suppose we want to solve the system of equations $Ax = b$. The problem may have no exact solution, precisely one solution, or there may be infinitely many solutions. If we minimize the function

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

we get a *least-squares* solution, generally, and an exact solution, whenever exact solutions exist. When the matrix $A$ is ill-conditioned, small changes in the vector $b$ can lead to large changes in the solution. When the vector $b$ comes from measured data, the entries of $b$ may include measurement errors, so that an exact solution of $Ax = b$ may be undesirable, even when such exact solutions exist; exact solutions may correspond to $x$ with unacceptably large norm, for example. In such cases, we may, instead, wish to minimize a function such as

$$\frac{1}{2}\|Ax - b\|_2^2 + \frac{\epsilon}{2}\|x - z\|_2^2, \tag{7.10}$$

for some vector $z$. If $z = 0$, the minimizing vector $x_\epsilon$ is then a *norm-constrained* least-squares solution. We then say that the least-squares problem has been *regularized*. In the limit, as $\epsilon \to 0$, these regularized solutions $x_\epsilon$ converge to the least-squares solution closest to $z$.

Suppose the system $Ax = b$ has infinitely many exact solutions. Our problem is to select one. Let us select $z$ that incorporates features of the desired solution, to the extent that we know them *a priori*. Then, as $\epsilon \to 0$, the vectors $x_\epsilon$ converge to the exact solution closest to $z$. For example, taking $z = 0$ leads to the *minimum-norm solution*.

**Minimizing Cross-Entropy**

In image processing, it is common to encounter systems $Px = y$ in which all the terms are non-negative. In such cases, it may be desirable to solve the system $Px = y$, approximately, perhaps, by minimizing the *cross-entropy* or *Kullback-Leibler distance*

$$KL(y, Px) = \sum_{i=1}^{I} \Big( y_i \log \frac{y_i}{(Px)_i} + (Px)_i - y_i \Big), \tag{7.11}$$

over vectors $x \geq 0$. When the vector $y$ is noisy, the resulting solution, viewed as an image, can be unacceptable. It is wise, therefore, to add a penalty term, such as $p(x) = \epsilon KL(z, x)$, where $z > 0$ is a prior estimate of the desired $x$ [113, 149, 114, 30].

A similar problem involves minimizing the function $KL(Px, y)$. Once again, noisy results can be avoided by including a penalty term, such as $p(x) = \epsilon KL(x, z)$ [30].

### The Lagrangian in Convex Programming

When there is a sensitivity vector $\lambda$ for the CP problem, minimizing $f(x)$ is equivalent to minimizing the Lagrangian,

$$f(x) + \sum_{i=1}^{I} \lambda_i g_i(x) = f(x) + p(x); \tag{7.12}$$

in this case, the addition of the second term, $p(x)$, serves to incorporate the constraints $g_i(x) \leq 0$ in the function to be minimized, turning a constrained minimization problem into an unconstrained one. The problem of minimizing the Lagrangian still remains, though. We may have to solve that problem using an iterative algorithm.

### Infimal Convolution

The *infimal convolution* of the functions $f$ and $g$ is defined as

$$(f \oplus g)(z) = \inf_x \Big\{ f(x) + g(z - x) \Big\}.$$

The *infimal deconvolution* of $f$ and $g$ is defined as

$$(f \ominus g)(z) = \sup_x \Big\{ f(z - x) - g(x) \Big\}.$$

Later we shall relate the infimal convolution and deconvolution to the Fenchel conjugate.

### Moreau's Proximity-Function Method

The Moreau envelope of the function $f$ is the function

$$m_f(z) = \inf_x \Big\{ f(x) + \frac{1}{2} \|x - z\|_2^2 \Big\}, \tag{7.13}$$

which is also the *infimal convolution* of the functions $f(x)$ and $\frac{1}{2}\|x\|_2^2$. It can be shown that the infimum is uniquely attained at the point denoted $x = \text{prox}_f z$ (see [136]). In similar fashion, we can define $m_{f^*} z$ and $\text{prox}_{f^*} z$, where $f^*(z)$ denotes the function conjugate to $f$.

**Proposition 7.1** *The infimum of $m_f(z)$, over all $z$, is the same as the infimum of $f(x)$, over all $x$.*

**Proof:** We have

$$\inf_z m_f(z) = \inf_z \inf_x \{f(x) + \frac{1}{2}\|x - z\|_2^2\}$$

$$= \inf_x \inf_z \{f(x) + \frac{1}{2}\|x - z\|_2^2\} = \inf_x \{f(x) + \frac{1}{2} \inf_z \|x - z\|_2^2\} = \inf_x f(x).$$

∎

The minimizers of $m_f(z)$ and $f(x)$ are the same, as well. Therefore, one way to use Moreau's method is to replace the original problem of minimizing the possibly non-smooth function $f(x)$ with the problem of minimizing the smooth function $m_f(z)$. Another way is to convert Moreau's method into a sequential minimization algorithm, replacing $z$ with $x^{k-1}$ and minimizing with respect to $x$ to get $x^k$. As we shall see, this leads to the proximal minimization algorithm to be discussed in a later chapter.

### 7.5.3 The Roles Penalty Functions Play

From the examples just surveyed, we can distinguish several distinct roles that penalty functions can play.

#### Impose Constraints

The first role is to penalize violations of constraints, as part of sequential minimization, or even to turn a constrained minimization into an equivalent unconstrained one: the Absolute-Value and Courant-Beltrami penalty functions penalize violations of the constraints $g_i(x) \leq 0$, while Quadratic-Loss penalty function penalizes violations of the constraints $g_i(x) = 0$. The augmented objective functions $f(x) + kp(x)$ now become part of a sequential unconstrained minimization method. It is sometimes possible for $f(x)$ and $f(x) + p(x)$ to have the same minimizers, or for constrained minimizers of $f(x)$ to be the same as unconstrained minimizers of $f(x) + p(x)$, as happens with the Lagrangian in the CP problem.

#### Regularization

The second role is regularization: in the least-squares problem, the main purpose for adding the norm-squared penalty function in Equation (7.10) is to reduce sensitivity to noise in the entries of the vector $b$. Also, regularization will usually turn a problem with multiple solutions into one with a unique solution.

**Incorporate Prior Information**

The third role is to incorporate prior information: when $Ax = b$ is underdetermined, using the penalty function $\epsilon\|x - z\|_2^2$ and letting $\epsilon \to 0$ encourages the solution to be close to the prior estimate $z$.

**Simplify Calculations**

A fourth role that penalty functions can play is to simplify calculation: in the case of cross-entropy minimization, adding the penalty functions $KL(z, x)$ and $KL(x, z)$ to the objective functions $KL(y, Px)$ and $KL(Px, y)$, respectively, regularizes the minimization problem. But, as we shall see later, the SMART algorithm minimizes $KL(Px, y)$ by using a sequential approach, in which each minimizer $x^k$ can be calculated in closed form.

**Sequential Unconstrained Minimization**

More generally, a fifth role for penalty functions is as part of sequential minimization. Here the goal is to replace one computationally difficult minimization with a sequence of simpler ones. Clearly, one reason for the difficulty can be that the original problem is constrained, and the sequential approach uses a series of unconstrained minimizations, penalizing violations of the constraints through the penalty function. However, there are other instances in which the sequential approach serves to simplify the calculations, not to remove constraints, but, perhaps, to replace a nondifferentiable objective function with a differentiable one, or a sequence of differentiable ones, as in Moreau's method.

## 7.6 Proximity-Function Minimization (I)

Let $f : R^J \to (-\infty, +\infty]$ be closed, proper, convex and differentiable. Let $h$ be a closed proper convex function, with effective domain $D$, that is differentiable on the nonempty open convex set int $D$. Assume that $f(x)$ is finite on $C = \overline{D}$ and attains its minimum value on $C$ at $\hat{x}$. The corresponding *Bregman distance* $D_h(x, z)$ is defined for $x$ in $D$ and $z$ in int $D$ by

$$D_h(x, z) = h(x) - h(z) - \langle \nabla h(z), x - z \rangle. \tag{7.14}$$

Note that $D_h(x, z) \geq 0$ always. If $h$ is essentially strictly convex, then $D_h(x, z) = 0$ implies that $x = z$. Our objective is to minimize $f(x)$ over $x$ in $C = \overline{D}$.

### 7.6.1 Proximal Minimization Algorithm

At the $k$th step of the *proximal minimization algorithm* (PMA) [38], we minimize the function

$$G_k(x) = f(x) + D_h(x, x^{k-1}), \tag{7.15}$$

to get $x^k$. The function

$$g_k(x) = D_h(x, x^{k-1}) \tag{7.16}$$

is nonnegative and $g_k(x^{k-1}) = 0$. We assume that each $x^k$ lies in int $D$.

### 7.6.2 The Method of Auslander and Teboulle

In [4] Auslander and Teboulle consider an iterative method similar to the PMA, in which, at the $k$th step, one minimizes the function

$$F_k(x) = f(x) + d(x, x^{k-1}) \tag{7.17}$$

to get $x^k$. Their distance $d(x, y)$ is not assumed to be a Bregman distance. Instead, they assume that the distance $d$ has an associated *induced proximal distance* $H(a, b) \geq 0$, finite for $a$ and $b$ in $D$, with $H(a, a) = 0$ and

$$\langle \nabla_1 d(b, a), c - b \rangle \leq H(c, a) - H(c, b), \tag{7.18}$$

for all $c$ in $D$. The notation $\nabla_1 d(x, y)$ denotes the gradient with respect to the vector variable $x$.

If $d = D_h$, that is, if $d$ is a Bregman distance, then from the equation

$$\langle \nabla_1 d(b, a), c - b \rangle = D_h(c, a) - D_h(c, b) - D_h(b, a) \tag{7.19}$$

we see that $D_h$ has $H = D_h$ for its associated induced proximal distance, so $D_h$ is *self-proximal*, in the terminology of [4].

## 7.7 The Simultaneous MART (SMART) (I)

Our next example is the simultaneous multiplicative algebraic reconstruction technique (SMART). For $a > 0$ and $b > 0$, the Kullback-Leibler distance, $KL(a, b)$, is defined as

$$KL(a, b) = a \log \frac{a}{b} + b - a. \tag{7.20}$$

In addition, $KL(0, 0) = 0$, $KL(a, 0) = +\infty$ and $KL(0, b) = b$. The KL distance is then extended to nonnegative vectors coordinate-wise.

### 7.7.1 The SMART Iteration

The SMART minimizes the function $f(x) = KL(Px, y)$, over nonnegative vectors $x$. Here $y$ is a vector with positive entries, and $P$ is a matrix with nonnegative entries, such that $s_j = \sum_{i=1}^{I} P_{ij} > 0$. Denote by $\mathcal{X}$ the set of all nonnegative $x$ for which the vector $Px$ has only positive entries.

Having found the vector $x^{k-1}$, the next vector in the SMART sequence is $x^k$, with entries given by

$$x_j^k = x_j^{k-1} \exp s_j^{-1}\Big(\sum_{i=1}^{I} P_{ij} \log(y_i/(Px^{k-1})_i)\Big). \tag{7.21}$$

### 7.7.2 The EMML Iteration

The EMML algorithm minimizes the function $f(x) = KL(y, Px)$, over nonnegative vectors $x$. Having found the vector $x^{k-1}$, the next vector in the EMML sequence is $x^k$, with entries given by

$$x_j^k = x_j^{k-1} s_j^{-1}\Big(\sum_{i=1}^{I} P_{ij}(y_i/(Px^{k-1})_i)\Big). \tag{7.22}$$

### 7.7.3 The EMML and the SMART as Alternating Minimization

In [30] the SMART was derived using the following alternating minimization approach.

For each $x \in \mathcal{X}$, let $r(x)$ and $q(x)$ be the $I$ by $J$ arrays with entries

$$r(x)_{ij} = x_j P_{ij} y_i/(Px)_i, \tag{7.23}$$

and

$$q(x)_{ij} = x_j P_{ij}. \tag{7.24}$$

In the iterative step of the SMART we get $x^k$ by minimizing the function

$$KL(q(x), r(x^{k-1})) = \sum_{i=1}^{I}\sum_{j=1}^{J} KL(q(x)_{ij}, r(x^{k-1})_{ij})$$

over $x \geq 0$. Note that $KL(Px, y) = KL(q(x), r(x))$.

Similarly, the iterative step of the EMML is to minimize the function $KL(r(x^{k-1}), q(x))$ to get $x = x^k$. Note that $KL(y, Px) = KL(r(x), q(x))$.

Now we establish the basic results for the SUMMA.

## 7.8 Convergence Theorems for SUMMA

At the $k$th step of the SUMMA we minimize the function $G_k(x)$ to get $x^k \in P$. In practice, of course, this minimization may need to be performed iteratively; we shall not address this issue here, and shall assume that $x^k$ can be computed. We make the following additional assumptions.

**Assumption 1:** The functions $g_k(x)$ are finite-valued on the subset $P$.

**Assumption 2:** The functions $g_k(x)$ satisfy the inequality in (7.2); that is,

$$0 \leq g_k(x) \leq G_{k-1}(x) - G_{k-1}(x^{k-1}),$$

for $k = 2, 3, ...$ and all $x \in P$. Consequently,

$$g_k(x^{k-1}) = 0.$$

**Assumption 3:** There is a real number $\alpha$ with

$$\alpha \leq f(x),$$

for all $x$ in $S$.

**Assumption 4:** Each $x^k$ is in $P$.

Using these assumptions, we can conclude several things about the sequence $\{x^k\}$.

**Proposition 7.2** *The sequence $\{f(x^k)\}$ is decreasing, and the sequence $\{g_k(x^k)\}$ converges to zero.*

**Proof:** We have

$$f(x^{k+1}) + g_{k+1}(x^{k+1}) = G_{k+1}(x^{k+1}) \leq G_{k+1}(x^k) = f(x^k) + g_{k+1}(x^k) = f(x^k).$$

Therefore,

$$f(x^k) - f(x^{k+1}) \geq g_{k+1}(x^{k+1}) \geq 0.$$

Since the sequence $\{f(x^k)\}$ is decreasing and bounded below by $d$, the difference sequence must converge to zero. Therefore, the sequence $\{g_k(x^k)\}$ converges to zero. ∎

**Theorem 7.1** *The sequence $\{f(x^k)\}$ converges to $d$.*

**Proof:** Suppose that there is $D > d$ with

$$f(x^k) \geq D,$$

for all $k$. Then there is $z$ in $P$ with

$$f(x^k) \geq D > f(z) \geq d,$$

for all $k$. From

$$g_{k+1}(z) \leq G_k(z) - G_k(x^k),$$

we have

$$g_k(z) - g_{k+1}(z) \geq f(x^k) + g_k(x^k) - f(z) \geq f(x^k) - f(z) \geq D - f(z) > 0,$$

for all $k$. This says that the nonnegative sequence $\{g_k(z)\}$ is decreasing, but that successive differences remain bounded away from zero, which cannot happen. ∎

**Definition 7.1** *A real-valued function $p(x)$ on $R^J$ has* bounded level sets *if, for all real $\gamma$, the level set $\{x|p(x) \leq \gamma\}$ is bounded.*

**Theorem 7.2** *Let $S$ be a complete metric space, $f(x)$ be a continuous function, $d > -\infty$, and the restriction of $f(x)$ to $x$ in $P$ have bounded level sets. Then the sequence $\{x^k\}$ is bounded, and $f(x^*) = d$, for any cluster point $x^* \in S$. If $\hat{x}$ is the unique minimizer of $f(x)$ for $x \in P$, then $x^* = \hat{x}$ and $\{x^k\} \to \hat{x}$.*

**Proof:** From the previous theorem we have $f(x^*) = d$, for all cluster points $x^*$. But, by uniqueness, $x^* = \hat{x}$, and so $\{x^k\} \to \hat{x}$. ∎

**Corollary 7.1** *Let $C \subseteq R^J$ be closed and convex. Let $f(x) : R^J \to R$ be closed, proper and convex. If $\hat{x}$ is the unique minimizer of $f(x)$ over $x \in C$, the sequence $\{x^k\}$ converges to $\hat{x}$.*

**Proof:** Let $\iota_C(x)$ be the indicator function of the set $C$, that is, $\iota_C(x) = 0$, for all $x$ in $C$, and $\iota_C(x) = +\infty$, otherwise. Then the function $g(x) = f(x) + \iota_C(x)$ is closed, proper and convex. If $\hat{x}$ is unique, then we have

$$\{x|f(x) + \iota_C(x) \leq f(\hat{x})\} = \{\hat{x}\}.$$

Therefore, one of the level sets of $g(x)$ is bounded and nonempty. It follows from Corollary 8.7.1 of [136] that every level set of $g(x)$ is bounded, so that the sequence $\{x^k\}$ is bounded. ∎

If $\hat{x}$ is not unique, we may still be able to prove convergence of the sequence $\{x^k\}$, for particular cases of SUMMA, as we shall see shortly.

## 7.9 Barrier-Function Methods (II)

We return now to the barrier-function methods, to show that they are particular cases of the SUMMA. The iterative step of the barrier-function method can be formulated as follows: minimize

$$f(x) + [(k-1)f(x) + b(x)] \tag{7.25}$$

to get $x^k$. Since, for $k = 2, 3, ...$, the function

$$(k-1)f(x) + b(x) \tag{7.26}$$

is minimized by $x^{k-1}$, the function

$$g_k(x) = (k-1)f(x) + b(x) - (k-1)f(x^{k-1}) - b(x^{k-1}) \tag{7.27}$$

is nonnegative, and $x^k$ minimizes the function

$$G_k(x) = f(x) + g_k(x). \tag{7.28}$$

From

$$G_k(x) = f(x) + (k-1)f(x) + b(x) - f(x^{k-1}) - (k-1)f(x^{k-1}) - b(x^{k-1}),$$

it follows that

$$G_k(x) - G_k(x^k) = kf(x) + b(x) - kf(x^k) - b(x^k) = g_{k+1}(x),$$

so that $g_{k+1}(x)$ satisfies the condition in (7.2). This shows that the barrier-function method is a particular case of SUMMA.

The goal is to minimize the objective function $f(x)$, over $x$ in the closed set $C = \overline{D}$, the closure of $D$. In the barrier-function method, we minimize

$$f(x) + \frac{1}{k}b(x) \tag{7.29}$$

over $x$ in $D$ to get $x^k$. Each $x^k$ lies within $D$, so the method is an interior-point algorithm. If the sequence $\{x^k\}$ converges, the limit vector $x^*$ will be in $C$ and $f(x^*) = f(\hat{x})$.

From the results for SUMMA, we conclude that $\{f(x^k)\}$ is decreasing to $f(\hat{x})$, and that $\{g_k(x^k)\}$ converges to zero. From the nonnegativity of $g_k(x^k)$ we have that

$$(k-1)(f(x^k) - f(x^{k-1})) \geq b(x^{k-1}) - b(x^k).$$

Since the sequence $\{f(x^k)\}$ is decreasing, the sequence $\{b(x^k)\}$ must be increasing, but might not be bounded above.

If $\hat{x}$ is unique, and $f(x)$ has bounded level sets, then it follows, from our discussion of SUMMA, that $\{x^k\} \rightarrow \hat{x}$. Suppose now that $\hat{x}$ is not known to be unique, but can be chosen in $D$, so that $G_k(\hat{x})$ is finite for each $k$. From

$$f(\hat{x}) + \frac{1}{k}b(\hat{x}) \geq f(x^k) + \frac{1}{k}b(x^k)$$

we have

$$\frac{1}{k}\Big(b(\hat{x}) - b(x^k)\Big) \geq f(x^k) - f(\hat{x}) \geq 0,$$

so that

$$b(\hat{x}) - b(x^k) \geq 0,$$

for all $k$. If either $f$ or $b$ has bounded level sets, then the sequence $\{x^k\}$ is bounded and has a cluster point, $x^*$ in $C$. It follows that $b(x^*) \leq b(\hat{x}) < +\infty$, so that $x^*$ is in $D$. If we assume that $f(x)$ is convex and $b(x)$ is strictly convex on $D$, then we can show that $x^*$ is unique in $D$, so that $x^* = \hat{x}$ and $\{x^k\} \rightarrow \hat{x}$.

To see this, assume, to the contrary, that there are two distinct cluster points $x^*$ and $x^{**}$ in $D$, with

$$\{x^{k_n}\} \rightarrow x^*,$$

and

$$\{x^{j_n}\} \rightarrow x^{**}.$$

Without loss of generality, we assume that

$$0 < k_n < j_n < k_{n+1},$$

for all $n$, so that

$$b(x^{k_n}) \leq b(x^{j_n}) \leq b(x^{k_{n+1}}).$$

Therefore,

$$b(x^*) = b(x^{**}) \leq b(\hat{x}).$$

From the strict convexity of $b(x)$ on the set $D$, and the convexity of $f(x)$, we conclude that, for $0 < \lambda < 1$ and $y = (1-\lambda)x^* + \lambda x^{**}$, we have $b(y) < b(x^*)$ and $f(y) \leq f(x^*)$. But, we must then have $f(y) = f(x^*)$. There must then be some $k_n$ such that

$$G_{k_n}(y) = f(y) + \frac{1}{k_n}b(y) < f(x_{k_n}) + \frac{1}{k_n}b(x_{k_n}) = G_{k_n}(x^{k_n}).$$

But, this is a contradiction. ∎

The following theorem summarizes what we have shown with regard to the barrier-function method.

**Theorem 7.3** *Let $f : R^J \rightarrow (-\infty, +\infty]$ be a continuous function. Let $b(x) : R^J \rightarrow (0, +\infty]$ be a continuous function, with effective domain the nonempty set $D$. Let $\hat{x}$ minimize $f(x)$ over all $x$ in $C = \overline{D}$. For each positive integer $k$, let $x^k$ minimize the function $f(x) + \frac{1}{k}b(x)$. Then the sequence $\{f(x^k)\}$ is monotonically decreasing to the limit $f(\hat{x})$, and the sequence $\{b(x^k)\}$ is increasing. If $\hat{x}$ is unique, and $f(x)$ has bounded level sets, then the sequence $\{x^k\}$ converges to $\hat{x}$. In particular, if $\hat{x}$ can be chosen in $D$, if either $f(x)$ or $b(x)$ has bounded level sets, if $f(x)$ is convex and if $b(x)$ is strictly convex on $D$, then $\hat{x}$ is unique in $D$ and $\{x^k\}$ converges to $\hat{x}$.*

At the $k$th step of the barrier method we must minimize the function $f(x) + \frac{1}{k}b(x)$. In practice, this must also be performed iteratively, with, say, the Newton-Raphson algorithm. It is important, therefore, that barrier functions be selected so that relatively few Newton-Raphson steps are needed to produce acceptable solutions to the main problem. For more on these issues see Renegar [134] and Nesterov and Nemirovski [130].

## 7.10 Penalty-Function Methods (II)

Once again, our objective is to find a sequence $\{x^k\}$ such that $\{f(x^k)\} \rightarrow d$. We select a penalty function $p(x)$ with $p(x) \geq 0$ and $p(x) = 0$ if and only if $x$ is in $P$. For $k = 1, 2, ...$, let $x^k$ be a minimizer of the function $f(x) + kp(x)$. As we shall see, we can formulate this penalty-function algorithm as a barrier-function iteration.

### 7.10.1 Penalty-Function Methods as Barrier-Function Methods

In order to relate penalty-function methods to barrier-function methods, we note that minimizing $T_k(x) = f(x) + kp(x)$ is equivalent to minimizing $p(x) + \frac{1}{k}f(x)$. This is the form of the barrier-function iteration, with $p(x)$ now in the role previously played by $f(x)$, and $f(x)$ now in the role previously played by $b(x)$. We are not concerned here with the effective domain of $f(x)$. Therefore, we can now mimic most, but not all, of what we did for barrier-function methods.

### 7.10.2 Basic Facts

**Lemma 7.1** *The sequence $\{T_k(x^k)\}$ is increasing, bounded above by $d$ and converges to some $\gamma \leq d$.*

**Proof:** We have

$$T_k(x^k) \leq T_k(x^{k+1}) \leq T_k(x^{k+1}) + p(x^{k+1}) = T_{k+1}(x^{k+1}).$$

Also, for any $z \in P$, and for each $k$, we have

$$f(z) = f(z) + kp(z) = T_k(z) \geq T_k(x^k);$$

therefore $d \geq \gamma$. ∎

**Lemma 7.2** *The sequence $\{p(x^k)\}$ is decreasing to zero, the sequence $\{f(x^k)\}$ is increasing and converging to some $\beta \leq d$.*

**Proof:** Since $x^k$ minimizes $T_k(x)$ and $x^{k+1}$ minimizes $T_{k+1}(x)$, we have

$$f(x^k) + kp(x^k) \leq f(x^{k+1}) + kp(x^{k+1}),$$

and

$$f(x^{k+1}) + (k+1)p(x^{k+1}) \leq f(x^k) + (k+1)p(x^k).$$

Consequently, we have

$$(k+1)[p(x^k) - p(x^{k+1})] \geq f(x^{k+1}) - f(x^k) \geq k[p(x^k) - p(x^{k+1})].$$

Therefore,

$$p(x^k) - p(x^{k+1}) \geq 0,$$

and

$$f(x^{k+1}) - f(x^k) \geq 0.$$

From

$$f(x^k) \leq f(x^k) + kp(x^k) = T_k(x^k) \leq \gamma \leq d,$$

it follows that the sequence $\{f(x^k)\}$ is increasing and converges to some $\beta \leq \gamma$. Since

$$\alpha + kp(x^k) \leq f(x^k) + kp(x^k) = T_k(x^k) \leq \gamma$$

for all $k$, we have $0 \leq kp(x^k) \leq \gamma - \alpha$. Therefore, the sequence $\{p(x^k)\}$ converges to zero. ∎

We want $\beta = d$. To obtain this result, it appears that we need to make more assumptions: we assume $S$ is a complete metric space, $P$ is closed in $S$, the functions $f$ and $p$ are continuous and $f$ has bounded level sets. From these assumptions, we are able to assert that the sequence $\{x^k\}$ is bounded, so that there is a convergent subsequence; let $\{x^{k_n}\} \to x^*$. It follows that $p(x^*) = 0$, so that $x^*$ is in $P$. Then

$$f(x^*) = f(x^*) + p(x^*) = \lim_{n \to +\infty} (f(x^{k_n}) + p(x^{k_n})) \leq \lim_{n \to +\infty} T_{k_n}(x^{k_n}) = \gamma \leq d.$$

But $x^* \in P$, so $f(x^*) \geq d$. Therefore, $f(x^*) = d$.

It may seem odd that we are trying to minimize $f(x)$ over the set $P$ using a sequence $\{x^k\}$ with $\{f(x^k)\}$ increasing, but remember that these $x^k$ are not in $P$.

## 7.11 Proximal Minimization Algorithms (II)

We show now that Assumption 3 holds, so that the PMA is a particular case of the SUMMA. We remind the reader that $f(x)$ is now assumed to be convex and differentiable, so that the Bregman distance $D_f(x, z)$ is defined and nonnegative, for all $x$ in $D$ and $z$ in int$D$.

**Lemma 7.3** *For each $k$ we have*

$$G_k(x) = G_k(x^k) + D_f(x, x^k) + D_h(x, x^k). \tag{7.30}$$

**Proof:** Since $x^k$ minimizes $G_k(x)$ within the set $D$, we have

$$0 = \nabla f(x^k) + \nabla h(x^k) - \nabla h(x^{k-1}). \tag{7.31}$$

Then

$$G_k(x) - G_k(x^k) = f(x) - f(x^k) + h(x) - h(x^k) - \langle \nabla h(x^{k-1}), x - x^k \rangle.$$

Now substitute, using Equation (7.31), and use the definition of Bregman distances. ∎

It follows from Lemma 7.3 that

$$G_k(x) - G_k(x^k) = g_{k+1}(x) + D_f(x, x^k),$$

so Assumption 3 holds.

From the discussion of the SUMMA we know that $\{f(x^k)\}$ is monotonically decreasing to $f(\hat{x})$. As we noted previously, if the sequence $\{x^k\}$ is bounded, and $\hat{x}$ is unique, we can conclude that $\{x^k\} \to \hat{x}$.

Suppose that $\hat{x}$ is not known to be unique, but can be chosen in $D$; this will be the case, of course, whenever $D$ is closed. Then $G_k(\hat{x})$ is finite for each $k$. From the definition of $G_k(x)$ we have

$$G_k(\hat{x}) = f(\hat{x}) + D_h(\hat{x}, x^{k-1}). \tag{7.32}$$

From Equation (7.30) we have

$$G_k(\hat{x}) = G_k(x^k) + D_f(\hat{x}, x^k) + D_h(\hat{x}, x^k), \tag{7.33}$$

so that

$$G_k(\hat{x}) = f(x^k) + D_h(x^k, x^{k-1}) + D_f(\hat{x}, x^k) + D_h(\hat{x}, x^k). \tag{7.34}$$

Therefore,

$$D_h(\hat{x}, x^{k-1}) - D_h(\hat{x}, x^k) =$$

$$f(x^k) - f(\hat{x}) + D_h(x^k, x^{k-1}) + D_f(\hat{x}, x^k). \tag{7.35}$$

It follows that the sequence $\{D_h(\hat{x}, x^k)\}$ is decreasing and that the sequence $\{D_f(\hat{x}, x^k)\}$ converges to 0. If either the function $f(x)$ or the function $D_h(\hat{x}, \cdot)$ has bounded level sets, then the sequence $\{x^k\}$ is bounded, has cluster points $x^*$ in $C$, and $f(x^*) = f(\hat{x})$, for every $x^*$. We now show that $\hat{x}$ in $D$ implies that $x^*$ is also in $D$, whenever $h$ is a Bregman -Legendre function.

Let $x^*$ be an arbitrary cluster point, with $\{x^{k_n}\} \rightarrow x^*$. If $\hat{x}$ is not in the interior of $D$, then, by Property B2 of Bregman-Legendre functions, we know that

$$D_h(x^*, x^{k_n}) \rightarrow 0,$$

so $x^*$ is in $D$. Then the sequence $\{D_h(x^*, x^k)\}$ is decreasing. Since a subsequence converges to zero, we have $\{D_h(x^*, x^k)\} \rightarrow 0$. From Property R5, we conclude that $\{x^k\} \rightarrow x^*$.

If $\hat{x}$ is in int $D$, but $x^*$ is not, then $\{D_h(\hat{x}, x^k)\} \rightarrow +\infty$, by Property R2. But, this is a contradiction; therefore $x^*$ is in $D$. Once again, we conclude that $\{x^k\} \rightarrow x^*$.

Now we summarize our results for the PMA. Let $f : R^J \rightarrow (-\infty, +\infty]$ be closed, proper, convex and differentiable. Let $h$ be a closed proper convex function, with effective domain $D$, that is differentiable on the nonempty open convex set int $D$. Assume that $f(x)$ is finite on $C = \overline{D}$ and attains its minimum value on $C$ at $\hat{x}$. For each positive integer $k$, let $x^k$ minimize the function $f(x) + D_h(x, x^{k-1})$. Assume that each $x^k$ is in the interior of $D$.

**Theorem 7.4** *If the restriction of $f(x)$ to $x$ in $C$ has bounded level sets and $\hat{x}$ is unique, and then the sequence $\{x^k\}$ converges to $\hat{x}$.*

**Theorem 7.5** *If $h(x)$ is a Bregman-Legendre function and $\hat{x}$ can be chosen in $D$, then $\{x^k\} \rightarrow x^*$, $x^*$ in $D$, with $f(x^*) = f(\hat{x})$.*

### 7.11.1 The Method of Auslander and Teboulle

The method of Auslander and Teboulle described in a previous section seems not to be a particular case of SUMMA. However, we can adapt the proof of Theorem 7.1 to prove the analogous result for their method. Once again, we assume that $f(\hat{x}) \leq f(x)$, for all $x$ in $C$.

**Theorem 7.6** *For $k = 2, 3, ...$, let $x^k$ minimize the function*

$$F_k(x) = f(x) + d(x, x^{k-1}).$$

*If the distance $d$ has an induced proximal distance $H$, then $\{f(x^k)\} \rightarrow f(\hat{x})$.*

**Proof:** First, we show that the sequence $\{f(x^k)\}$ is decreasing. We have

$$f(x^{k-1}) = F_k(x^{k-1}) \geq F_k(x^k) = f(x^k) + d(x^k, x^{k-1}),$$

from which we conclude that the sequence $\{f(x^k)\}$ is decreasing and the sequence $\{d(x^k, x^{k-1})\}$ converges to zero.

Now suppose that

$$f(x^k) \geq f(\hat{x}) + \delta,$$

for some $\delta > 0$ and all $k$. Since $\hat{x}$ is in $C$, there is $z$ in $D$ with

$$f(x^k) \geq f(z) + \frac{\delta}{2},$$

for all $k$. Since $x^k$ minimizes $F_k(x)$, it follows that

$$0 = \nabla f(x^k) + \nabla_1 d(x^k, x^{k-1}).$$

Using the convexity of the function $f(x)$ and the fact that $H$ is an induced proximal distance, we have

$$0 < \frac{\delta}{2} \leq f(x^k) - f(z) \leq \langle -\nabla f(x^k), z - x^k \rangle =$$

$$\langle \nabla_1 d(x^k, x^{k-1}), z - x^k \rangle \leq H(z, x^{k-1}) - H(z, x^k).$$

Therefore, the nonnegative sequence $\{H(z, x^k)\}$ is decreasing, but its successive differences remain bounded below by $\frac{\delta}{2}$, which is a contradiction. ∎

It is interesting to note that the Auslander-Teboulle approach places a restriction on the function $d(x, y)$, the existence of the induced proximal distance $H$, that is unrelated to the objective function $f(x)$, but this condition is helpful only for convex $f(x)$. In contrast, the SUMMA approach requires that

$$0 \leq g_{k+1}(x) \leq G_k(x) - G_k(x^k),$$

which involves the $f(x)$ being minimized, but does not require that this $f(x)$ be convex.

## 7.12 The Simultaneous MART (II)

It follows from the identities established in [30] that the SMART can also be formulated as a particular case of the SUMMA.

### 7.12.1 The SMART as a Case of SUMMA

We show now that the SMART is a particular case of the SUMMA. The following lemma is helpful in that regard.

**Lemma 7.4** *For any non-negative vectors $x$ and $z$, with $z_+ = \sum_{j=1}^J z_j > 0$, we have*

$$KL(x, z) = KL(x_+, z_+) + KL(x, \frac{x_+}{z_+}z). \tag{7.36}$$

For notational convenience, we assume, for the remainder of this chapter, that $s_j = 1$ for all $j$. From the identities established for the SMART in [30], we know that the iterative step of SMART can be expressed as follows: minimize the function

$$G_k(x) = KL(Px, y) + KL(x, x^{k-1}) - KL(Px, Px^{k-1}) \tag{7.37}$$

to get $x^k$. According to Lemma 7.4, the quantity

$$g_k(x) = KL(x, x^{k-1}) - KL(Px, Px^{k-1})$$

is nonnegative, since $s_j = 1$. The $g_k(x)$ are defined for all nonnegative $x$; that is, the set $D$ is the closed nonnegative orthant in $R^J$. Each $x^k$ is a positive vector.

It was shown in [30] that

$$G_k(x) = G_k(x^k) + KL(x, x^k), \tag{7.38}$$

from which it follows immediately that Assumption 3 holds for the SMART.

Because the SMART is a particular case of the SUMMA, we know that the sequence $\{f(x^k)\}$ is monotonically decreasing to $f(\hat{x})$. It was shown in [30] that if $y = Px$ has no nonnegative solution and the matrix $P$ and every submatrix obtained from $P$ by removing columns has full rank, then $\hat{x}$ is unique; in that case, the sequence $\{x^k\}$ converges to $\hat{x}$. As we shall see, the SMART sequence always converges to a nonnegative minimizer of $f(x)$. To establish this, we reformulate the SMART as a particular case of the PMA.

### 7.12.2 The SMART as a Case of the PMA

We take $F(x)$ to be the function

$$F(x) = \sum_{j=1}^J x_j \log x_j. \tag{7.39}$$

Then

$$D_F(x, z) = KL(x, z). \tag{7.40}$$

For nonnegative $x$ and $z$ in $\mathcal{X}$, we have

$$D_f(x, z) = KL(Px, Pz). \tag{7.41}$$

**Lemma 7.5** $D_F(x,z) \geq D_f(x,z)$.

**Proof:** We have

$$D_F(x,z) \geq \sum_{j=1}^{J} KL(x_j, z_j) \geq \sum_{j=1}^{J} \sum_{i=1}^{I} KL(P_{ij}x_j, P_{ij}z_j)$$
$$\geq \sum_{i=1}^{I} KL((Px)_i, (Pz)_i) = KL(Px, Pz). \tag{7.42}$$

∎

Then we let $h(x) = F(x) - f(x)$; then $D_h(x,z) \geq 0$ for nonnegative $x$ and $z$ in $\mathcal{X}$. The iterative step of the SMART is to minimize the function

$$f(x) + D_h(x, x^{k-1}). \tag{7.43}$$

So the SMART is a particular case of the PMA.

The function $h(x) = F(x) - f(x)$ is finite on $D$ the nonnegative orthant of $R^J$, and differentiable on the interior, so $C = D$ is closed in this example. Consequently, $\hat{x}$ is necessarily in $D$. From our earlier discussion of the PMA, we can conclude that the sequence $\{D_h(\hat{x}, x^k)\}$ is decreasing and the sequence $\{D_f(\hat{x}, x^k)\} \to 0$. Since the function $KL(\hat{x}, \cdot)$ has bounded level sets, the sequence $\{x^k\}$ is bounded, and $f(x^*) = f(\hat{x})$, for every cluster point. Therefore, the sequence $\{D_h(x^*, x^k)\}$ is decreasing. Since a subsequence converges to zero, the entire sequence converges to zero. The convergence of $\{x^k\}$ to $x^*$ follows from basic properties of the KL distance.

From the fact that $\{D_f(\hat{x}, x^k)\} \to 0$, we conclude that $P\hat{x} = Px^*$. Equation (7.35) now tells us that the difference $D_h(\hat{x}, x^{k-1}) - D_h(\hat{x}, x^k)$ depends on only on $P\hat{x}$, and not directly on $\hat{x}$. Therefore, the difference $D_h(\hat{x}, x^0) - D_h(\hat{x}, x^*)$ also depends only on $P\hat{x}$ and not directly on $\hat{x}$. Minimizing $D_h(\hat{x}, x^0)$ over nonnegative minimizers $\hat{x}$ of $f(x)$ is therefore equivalent to minimizing $D_h(\hat{x}, x^*)$ over the same vectors. But the solution to the latter problem is obviously $\hat{x} = x^*$. Thus we have shown that the limit of the SMART is the nonnegative minimizer of $KL(Px, y)$ for which the distance $KL(x, x^0)$ is minimized.

The following theorem summarizes the situation with regard to the SMART.

**Theorem 7.7** *In the consistent case the SMART converges to the unique nonnegative solution of $y = Px$ 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 $KL(Px, y)$ for which $\sum_{j=1}^{J} s_j KL(x_j, x_j^0)$ is minimized; if $P$ and every matrix derived from $P$ by deleting columns has full rank then there is a unique nonnegative minimizer of $KL(Px, y)$ and at most $I - 1$ of its entries are nonzero.*

### 7.12.3 The EMML Algorithm

The *expectation maximization maximum likelihood* (EMML) algorithm minimizes the function $f(x) = KL(y, Px)$ over $x$ in $\mathcal{X}$. In [44] the EMML algorithm and the SMART are developed in tandem to reveal how closely related these two methods are. There, the EMML algorithm is derived using alternating minimization, in which the vector $x^k$ is the one for which the function $KL(r(x^{k-1}), q(x))$ is minimized. When we try to put the EMML into the framework of SUMMA, we find that $x^k$ minimizes the function

$$G_k(x) = f(x) + KL(r(x^{k-1}), r(x)), \tag{7.44}$$

over all positive vectors $x$. However, the functions

$$g_k(x) = KL(r(x^{k-1}), r(x)) \tag{7.45}$$

appear not to satisfy the condition in (7.2).

It turns out, however, that the EMML algorithm is a particular case of SUMMA, but not in the most obvious way. In a later chapter on alternating minimization (alt min) we show that any alt min algorithm for which the three- and four-point properties hold is a SUMMA algorithm. Since the EMML is such an alt min algorithm, it must be the case that EMML is a SUMMA algorithm. The details are in the later chapter.

In the next section we present a variant of the SMART, designed to incorporate bounds on the entries of the vector $x$.

## 7.13 Minimizing $KL(Px, y)$ with upper and lower bounds on the vector $x$

Let $a_j < b_j$, for each $j$. Let $\mathcal{X}_{ab}$ be the set of all vectors $x$ such that $a_j \leq x_j \leq b_j$, for each $j$. Now, we seek to minimize $f(x) = KL(Px, y)$, over all vectors $x$ in $\mathcal{X} \cap \mathcal{X}_{ab}$. We let

$$F(x) = \sum_{j=1}^{J} \Big( (x_j - a_j)\log(x_j - a_j) + (b_j - x_j)\log(b_j - x_j) \Big). \tag{7.46}$$

Then we have

$$D_F(x, z) = \sum_{j=1}^{J} \Big( KL(x_j - a_j, z_j - a_j) + KL(b_j - x_j, b_j - z_j) \Big), \tag{7.47}$$

and, as before,

$$D_f(x, z) = KL(Px, Pz). \tag{7.48}$$

**Lemma 7.6** *For any $c > 0$, with $a \geq c$ and $b \geq c$, we have $KL(a-c, b-c) \geq KL(a, b)$.*

**Proof:** Let $g(c) = KL(a-c, b-c)$ and differentiate with respect to $c$, to obtain

$$g'(c) = \frac{a-c}{b-c} - 1 - \log(\frac{a-c}{b-c}) \geq 0. \tag{7.49}$$

We see then that the function $g(c)$ is increasing with $c$. ∎

As a corollary of Lemma 7.6, we have

**Lemma 7.7** *Let $a = (a_1, ..., a_J)^T$, and $x$ and $z$ in $\mathcal{X}$ with $(Px)_i \geq (Pa)_i$, $(Pz)_i \geq (Pa)_i$, for each $i$. Then $KL(Px, Pz) \leq KL(Px - Pa, Pz - Pa)$.*

**Lemma 7.8** *$D_F(x, z) \geq D_f(x, z)$.*

**Proof:** We can easily show that

$$D_F(x, z) \geq KL(Px - Pa, Pz - Pa) + KL(Pb - Px, Pb - Pz),$$

along the lines used previously. Then, from Lemma 7.7, we have

$$KL(Px - Pa, Pz - Pa) \geq KL(Px, Pz) = D_f(x, z).$$

∎

Once again, we let $h(x) = F(x) - f(x)$, which is finite on the closed convex set $\mathcal{X} \cap \mathcal{X}_{ab}$. At the $k$th step of this algorithm we minimize the function

$$f(x) + D_h(x, x^{k-1}) \tag{7.50}$$

to get $x^k$.

Solving for $x_j^k$, we obtain

$$x_j^{k+1} = \alpha_j^k a_j + (1 - \alpha_j^k) b_j, \tag{7.51}$$

where

$$(\alpha_j^k)^{-1} = 1 + \Big(\frac{x_j^{k-1} - a_j}{b_j - x_j^{k-1}}\Big) \exp\Big(\sum_{i=1}^{I} P_{ij} \log(y_i / (Px^{k-1})_i)\Big). \tag{7.52}$$

Since the restriction of $f(x)$ to $\mathcal{X} \cap \mathcal{X}_{ab}$ has bounded level sets, the sequence $\{x^k\}$ is bounded and has cluster points. If $\hat{x}$ is unique, then $\{x^k\} \to \hat{x}$. This algorithm is closely related to those presented in [36].

## 7.14 Computation

As we noted previously, we do not address computational issues in any detail in this chapter. Nevertheless, it cannot be ignored that both Equation (7.21) for the SMART and Equations (7.51) and (7.52) for the generalized SMART provide easily calculated iterates, in contrast to other examples of SUMMA. At the same time, showing that these two algorithms are particular cases of SUMMA requires the introduction of functions $G_k(x)$ that appear to be quite ad hoc. The purpose of this section is to motivate these choices of $G_k(x)$ and to indicate how other analogous computationally tractable SUMMA iterative schemes may be derived.

### 7.14.1 Landweber's Algorithm

Suppose that $A$ is a real $I$ by $J$ matrix and we wish to obtain a least-squares solution $\hat{x}$ of $Ax = b$ by minimizing the function

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

We know that

$$(A^T A)\hat{x} = A^T b, \tag{7.53}$$

so, in a sense, the problem is solved. However, in many applications, the dimensions $I$ and $J$ are quite large, perhaps in the tens of thousands, as in some image reconstruction problems. Solving Equation (7.53), and even calculating $A^T A$, can be prohibitively expensive. In such cases, we turn to iterative methods, not necessarily to incorporate constraints on $x$, but to facilitate calculation. Landweber's algorithm is one such iterative method for calculating a least-squares solution.

The iterative step of Landweber's algorithm is

$$x^k = x^{k-1} - \gamma A^T(Ax^{k-1} - b). \tag{7.54}$$

The sequence $\{x^k\}$ converges to the least-squares solution closest to $x^0$, for any choice of $\gamma$ in the interval $(0, 2/\rho(A^T A))$, where $\rho(A^T A)$, the spectral radius of $A^T A$, is its largest eigenvalue; this is a consequence of the Krasnoselskii-Mann Theorem.

It is easy to verify that the $x^k$ given by Equation (7.54) is the minimizer of the function

$$G_k(x) = \frac{1}{2}\|Ax - b\|_2^2 + \frac{1}{2\gamma}\|x - x^{k-1}\|_2^2 - \frac{1}{2}\|Ax - Ax^{k-1}\|_2^2, \tag{7.55}$$

that, for $\gamma$ in the interval $(0, 1/\rho(A^T A))$, the iteration in Equation (7.54) is a particular case of SUMMA, and

$$G_k(x) - G_k(x^k) = \frac{1}{2\gamma}\|x - x^k\|_2^2.$$

The similarity between the $G_k(x)$ in Equation (7.55) and that in Equation (7.37) is not accidental and both are particular cases of a more general iterative scheme involving proximal minimization.

### 7.14.2 Extending the PMA

The proximal minimization algorithm (PMA) requires us to minimize the function $G_k(x)$ given by Equation (7.15) to get $x^k$. How $x^k$ may be calculated was not addressed previously. Suppose, instead of minimizing $G_k(x)$ in Equation (7.15), we minimize

$$G_k(x) = f(x) + D_h(x, x^{k-1}) - D_f(x, x^{k-1}) = f(x) + g_k(x), \tag{7.56}$$

with the understanding that $f(x)$ is convex and

$$D_h(x, z) - D_f(x, z) \geq 0,$$

for all appropriate $x$ and $z$. The next iterate $x^k$ satisfies the equation

$$0 = \nabla h(x^k) - \nabla h(x^{k-1}) + \nabla f(x^{k-1}), \tag{7.57}$$

so that

$$\nabla h(x^k) = \nabla h(x^{k-1}) - \nabla f(x^{k-1}). \tag{7.58}$$

This iterative scheme is the *interior-point algorithm* (IPA) presented in [38]. If the function $h(x)$ is chosen carefully, then we can solve for $x^k$ easily. The Landweber algorithm, the SMART, and the generalized SMART are all particular cases of this IPA.

Using Lemma 7.3, we can show that

$$G_k(x) - G_k(x^k) = D_h(x, x^k) \geq g_{k+1}(x), \tag{7.59}$$

for all appropriate $x$, so that the IPA is a particular case of SUMMA. We consider now several other examples.

If we let $h(x) = \frac{1}{2\gamma}\|x\|_2^2$ in Equation (7.56), the iteration becomes

$$x^k = x^{k-1} - \gamma \nabla f(x^{k-1}). \tag{7.60}$$

If, for example, the operator $\nabla f$ is $L$-Lipschitz continuous, that is,

$$\|\nabla f(x) - \nabla f(z)\|_2 \leq L\|x - z\|_2,$$

then, for any $\gamma$ in the interval $(0, 1/2L)$, we have

$$\frac{1}{2\gamma}\|x - z\|_2^2 \geq L\|x - z\|_2^2 \geq \langle \nabla f(x) - \nabla f(z), x - z \rangle$$

$$= D_f(x,z) + D_f(z,x) \geq D_f(x,z).$$

Therefore, this iteration is a particular case of SUMMA. It should be noted that, in this case, the Krasnoselskii-Mann Theorem gives convergence for any $\gamma$ in the interval $(0, 2/L)$.

Finally, we consider what happens if we replace the Euclidean norm with that induced by the local geometry derived from $f$ itself. More specifically, let us take

$$h(x) = \frac{1}{2}x^T\nabla^2 f(x^{k-1})x,$$

so that

$$D_h(x, x^{k-1}) = \frac{1}{2}(x - x^{k-1})^T\nabla^2 f(x^{k-1})(x - x^{k-1}).$$

Then the IPA iterate $x^k$ becomes

$$x^k = x^{k-1} - \nabla^2 f(x^{k-1})^{-1}\nabla f(x^{k-1}), \tag{7.61}$$

which is the Newton-Raphson iteration. Using the SUMMA framework to study the Newton-Raphson method is work in progress.

Algorithms such as Landweber's and SMART can be slow to converge. It is known that convergence can often be accelerated using incremental gradient (partial gradient, block-iterative, ordered-subset) methods. Using the SUMMA framework to study such incremental gradient methods as the algebraic reconstruction technique (ART), its multiplicative version (MART), and other block-iterative methods is also the subject of on-going work.

## 7.15 Connections with Karmarkar's Method

As related by Margaret Wright in [152], a revolution in mathematical programming took place around 1984. In that year Narenda Karmarkar discovered the first efficient polynomial-time algorithm for the linear programming problem [107]. Khachian's earlier polynomial-time algorithm for LP was too slow and conventional wisdom prior to 1984 was that the simplex method was "the only game in town" . It was known that, for certain peculiar LP problems, the complexity of the simplex method grew exponentially with the size of the problem, and obtaining a polynomial-time method for LP had been a goal for quite a while. However, for most problems, the popular simplex method was more than adequate. Soon after Karmarkar's result was made known, others discovered that there was a close connection between this method and earlier barrier-function approaches in nonlinear programming [89]. This discovery not only revived barrier-function methods, but established a link between linear and nonlinear programming, two areas that had historically been treated separately.

The primary LP problem in standard form is to minimize $c^T x$, subject to the conditions $Ax = b$ and $x \geq 0$. The barrier-function approach is to use a logarithmic barrier to enforce the condition $x \geq 0$, and then to use the primal-dual approach to maintain the condition $Ax = b$. The function to be minimized, subject to $Ax = b$, is then

$$c^T x - \mu \sum_{j=1}^{J} \log x_j,$$

where $\mu > 0$ is the *barrier parameter*. When this minimization is performed using the primal-dual method, and the NR iteration is begun at a feasible $x^0$, each subsequent $x^k$ satisfies $Ax^k = b$. The limit of the NR iteration is $x_\mu$. Under reasonable conditions, $x_\mu$ will converge to the solution of the LP problem, as $\mu \to 0$. This interior-point approach to solving the LP problem is essentially equivalent to Karmarkar's method.

# Chapter 8

# The Forward-Backward Splitting Algorithm

## 8.1 The Context

In my course on applied linear algebra I presented the basic theory of averaged operators, in order to get to the projected gradient-descent method, the $CQ$ algorithm and projected Landweber. Over spring break in 2012 I began to consider if it was possible to bypass this theory and obtain a more elementary proof. The problem lies with showing that, whenever the gradient of a convex function is non-expansive, it is firmly non-expansive, and therefore averaged. This is a non-trivial result and I am forced to skip its proof in the class. It is also non-trivial to prove that the product of averaged operators is averaged. I wanted to see how far the SUMMA approach could be pushed to prove convergence of these algorithms.

My first result along these lines was that convergence of the gradient-descent method could be established using SUMMA. I did not see how to include projection, so I wrote up a short note on this and sent it to my list of colleagues and to JOTA. Within a day or two, I realized that projection can be included, simply by minimizing $G_k(x)$ over $x$ in $C$. This time, I just sent it to my friends, not to JOTA, with the promise not to become a nuisance. Unfortunately, I was not able to keep my word.

The following day I discovered how to apply the same SUMMA approach to prove convergence of the forward-backward splitting method of Combettes and Wajs. The pieces of the puzzle just fell into place very nicely, which is what tends to happen when you get the right idea in the first place. I wrote up a longer paper, the content of this chapter, and sent it to JOTA. In a comment to the editor, I said that I had no objection if they chose to consider the first note withdrawn and replace it with the

recent one.

## 8.2 Forward-Backward Splitting

Let $f : R^J \to R$ be convex. For each $z \in R^J$ the function

$$m_f(z) = \min_x \{f(x) + \frac{1}{2}\|x - z\|_2^2\}$$

is minimized by $x = \text{prox}_f(z)$. Moreau's proximity operator $\text{prox}_f$ extends the notion of orthogonal projection onto a closed convex set [124, 125, 126]. Proximity operators are also firmly non-expansive [67]. We have $x = \text{prox}_f(z)$ if and only if $z - x \in \partial f(x)$, where the set $\partial f(x)$ is the sub-differential of $f$ at $x$, given by

$$\partial f(x) = \{u | \langle u, y - x\rangle \leq f(y) - f(x), \text{for all } y\}.$$

Our objective here is to provide an elementary proof of convergence for the *forward-backward splitting* (FBS) algorithm; a detailed discussion of this algorithm and its history is given by Combettes and Wajs in [67].

**Theorem 8.1** *Let $f : R^J \to R$ be convex, with $f = f_1 + f_2$, both convex, $f_2$ differentiable, and $\nabla f_2$ L-Lipschitz. For $0 < \gamma < \frac{1}{L}$, let*

$$x^k = \text{prox}_{\gamma f_1}\Big(x^{k-1} - \gamma \nabla f_2(x^{k-1})\Big). \tag{8.1}$$

*The sequence $\{x^k\}$ converges to a minimizer of the function $f(x)$, whenever such minimizers exist.*

Any fixed point of the iteration minimizes the function $f(x)$. Because proximity operators are firmly non-expansive, and therefore averaged, it is a consequence of the Krasnoselskii-Mann Theorem [110, 119] 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 results that, because the operator $\frac{1}{L}\nabla f_2$ is non-expansive, it is firmly non-expansive, and that the product of averaged operators is averaged.

Several applications of the theorem are given, including the proof of convergence of two interior-point algorithms for minimizing $f(x)$ over $x$ with $Ax = b$.

## 8.3 Sequential Unconstrained Optimization

Sequential unconstrained optimization algorithms can be used to minimize a function $f : R^J \to (-\infty, \infty]$ over a (not necessarily proper) subset $C$

of $R^J$ [85]. At the $k$th step of a *sequential unconstrained minimization* method we obtain $x^k$ by minimizing the function

$$G_k(x) = f(x) + g_k(x), \tag{8.2}$$

where the auxiliary function $g_k(x)$ is appropriately chosen. If $C$ is a proper subset of $R^J$ we may force $g_k(x) = +\infty$ for $x$ not in $C$, as in the barrier-function methods; then each $x^k$ will lie in $C$. The objective is then to select the $g_k(x)$ so that the sequence $\{x^k\}$ converges to a solution of the problem, or failing that, at least to have the sequence $\{f(x^k)\}$ converging to the infimum of $f(x)$ over $x$ in $C$.

Our main focus in this paper is the use of sequential unconstrained optimization algorithms to obtain iterative methods in which each iterate can be obtained in closed form. Now the $g_k(x)$ are selected not to impose a constraint, but to facilitate computation.

## 8.4 SUMMA

In [46] we presented a particular class of sequential unconstrained minimization methods called SUMMA. As we showed in that paper, this class is broad enough to contain barrier-function methods, proximal minimization methods, and the simultaneous multiplicative algebraic reconstruction technique (SMART). By reformulating the problem, the penalty-function methods can also be shown to be members of the SUMMA class. Any alternating minimization (AM) problem with the five-point property [70] can be reformulated as a SUMMA problem; therefore the *expectation maximization maximum likelihood* (EMML) algorithm for Poisson data, which is such an AM algorithm, must also be a SUMMA algorithm.

For a method to be in the SUMMA class we require that $x^k \in C$ for each $k$ and that each auxiliary function $g_k(x)$ satisfy the inequality

$$0 \leq g_k(x) \leq G_{k-1}(x) - G_{k-1}(x^{k-1}), \tag{8.3}$$

for all $x$. Note that it follows that $g_k(x^{k-1}) = 0$, for all $k$. For this note we require that $f(x)$ be convex and differentiable, and that the gradient operator, $\nabla f$, be $L$-Lipschitz.

We assume, throughout this section, that the inequality in (8.3) holds for each $k$. We also assume that $\inf_{x \in C} f(x) = b > -\infty$. The next two results are taken from [46].

**Proposition 8.1** *The sequence $\{f(x^k)\}$ is non-increasing and the sequence $\{g_k(x^k)\}$ converges to zero.*

**Proof:** We have

$$f(x^{k+1}) + g_{k+1}(x^{k+1}) = G_{k+1}(x^{k+1}) \leq G_{k+1}(x^k) = f(x^k). \tag{8.4}$$

∎

**Theorem 8.2** *The sequence* $\{f(x^k)\}$ *converges to* $b$.

**Proof:** Suppose that there is $\delta > 0$ such that $f(x^k) \geq b+2\delta$, for all $k$. Then there is $z \in C$ such that $f(x^k) \geq f(z) + \delta$, for all $k$. From the inequality in (8.3) we have

$$g_k(z) - g_{k+1}(z) \geq f(x^k) + g_k(x^k) - f(z) \geq f(x^k) - f(z) \geq \delta, \qquad (8.5)$$

for all $k$. But this cannot happen; the successive differences of a non-increasing sequence of non-negative terms must converge to zero. ∎

## 8.5 Convergence of the FBS algorithm

For each $k = 1, 2, ...$ let

$$G_k(x) = f(x) + \frac{1}{2\gamma}\|x - x^{k-1}\|_2^2 - D_{f_2}(x, x^{k-1}), \qquad (8.6)$$

where

$$D_{f_2}(x, x^{k-1}) = f_2(x) - f_2(x^{k-1}) - \langle \nabla f_2(x^{k-1}), x - x^{k-1} \rangle. \qquad (8.7)$$

Since $f_2(x)$ is convex, $D_{f_2}(x, y) \geq 0$ for all $x$ and $y$ and is the Bregman distance formed from the function $f_2$ [15].

**Lemma 8.1** *The* $x^k$ *that minimizes* $G_k(x)$ *over* $x$ *is given by Equation (8.1).*

**Proof:** Since $x^k$ minimizes $G_k(x)$ we know that

$$0 \in \nabla f_2(x^k) + \frac{1}{\gamma}(x^k - x^{k-1}) - \nabla f_2(x^k) + \nabla f_2(x^{k-1}) + \partial f_1(x^k).$$

Therefore,

$$\left(x^{k-1} - \gamma \nabla f_2(x^{k-1})\right) - x^k \in \partial \gamma f_1(x^k).$$

Consequently,

$$x^k = \text{prox}_{\gamma f_1}(x^{k-1} - \gamma \nabla f_2(x^{k-1})).$$

∎

The auxiliary function

$$g_k(x) = \frac{1}{2\gamma}\|x - x^{k-1}\|_2^2 - D_{f_2}(x, x^{k-1}) \qquad (8.8)$$

can be rewritten as

$$g_k(x) = D_h(x, x^{k-1}), \tag{8.9}$$

where

$$h(x) = \frac{1}{2\gamma}\|x\|_2^2 - f_2(x). \tag{8.10}$$

Therefore, $g_k(x) \geq 0$ whenever $h(x)$ is a convex function.

We know that $h(x)$ is convex if and only if

$$\langle \nabla h(x) - \nabla h(y), x - y \rangle \geq 0, \tag{8.11}$$

for all $x$ and $y$. This is equivalent to

$$\frac{1}{\gamma}\|x - y\|_2^2 - \langle \nabla f_2(x) - \nabla f_2(y), x - y \rangle \geq 0. \tag{8.12}$$

Since $\nabla f_2$ is $L$-Lipschitz, the inequality (8.12) holds whenever $0 < \gamma < \frac{1}{L}$.

A relatively simple calculation shows that

$$G_k(x) - G_k(x^k) = \frac{1}{2\gamma}\|x - x^k\|_2^2 + \Big(f_1(x) - f_1(x^k) - \langle (x^{k-1} - \gamma \nabla f_2(x^{k-1})) - x^k, x - x^k \rangle\Big). \tag{8.13}$$

Since

$$(x^{k-1} - \gamma \nabla f_2(x^{k-1})) - x^k \in \partial \gamma f_1(x^k),$$

it follows that

$$\Big(f_1(x) - f_1(x^k) - \langle (x^{k-1} - \gamma \nabla f_2(x^{k-1})) - x^k, x - x^k \rangle\Big) \geq 0.$$

Therefore,

$$G_k(x) - G_k(x^k) \geq \frac{1}{2\gamma}\|x - x^k\|_2^2 \geq g_{k+1}(x). \tag{8.14}$$

Therefore, the inequality in (8.3) holds and the iteration fits into the SUMMA class.

Now let $\hat{x}$ minimize $f(x)$ over all $x$. Then

$$G_k(\hat{x}) - G_k(x^k) = f(\hat{x}) + g_k(\hat{x}) - f(x^k) - g_k(x^k)$$

$$\leq f(\hat{x}) + G_{k-1}(\hat{x}) - G_{k-1}(x^{k-1}) - f(x^k) - g_k(x^k),$$

so that

$$\Big(G_{k-1}(\hat{x}) - G_{k-1}(x^{k-1})\Big) - \Big(G_k(\hat{c}) - G_k(x^k)\Big) \geq f(x^k) - f(\hat{x}) + g_k(x^k) \geq 0.$$

Therefore, the sequence $\{G_k(\hat{x}) - G_k(x^k)\}$ is decreasing and the sequences $\{g_k(x^k)\}$ and $\{f(x^k) - f(\hat{x})\}$ converge to zero.

From

$$G_k(\hat{x}) - G_k(x^k) \geq \frac{1}{2\gamma}\|\hat{x} - x^k\|_2^2,$$

it follows that the sequence $\{x^k\}$ is bounded and that a subsequence converges to some $x^* \in C$ with $f(x^*) = f(\hat{x})$.

Replacing the generic $\hat{x}$ with $x^*$, we find that $\{G_k(x^*) - G_k(x^k)\}$ is decreasing. By Equation (8.13), it therefore converges to the limit

$$\frac{1}{2\gamma}\|x^* - x^*\|_2^2 + \frac{1}{\gamma}\langle(\text{prox}_{\gamma f_1} - I)(x^* - \gamma\nabla f(x^*)), x^* - \text{prox}_{\gamma f_1}(x^* - \gamma\nabla f(x^*))\rangle = 0.$$

From the inequality in (8.14), we conclude that the sequence $\{\|x^* - x^k\|_2^2\}$ converges to zero, and so $\{x^k\}$ converges to $x^*$. This completes the proof of the theorem.

## 8.6 Some Examples

We present some examples to illustrate the application of the convergence theorem.

### 8.6.1 Projected Gradient Descent

Let $C$ be a non-empty, closed convex subset of $R^J$ and $f_1(x) = \iota_C(x)$, the function that is $+\infty$ for $x$ not in $C$ and zero for $x$ in $C$. Then $\iota_C(x)$ is convex, but not differentiable. We have $\text{prox}_{\gamma f_1} = P_C$, the orthogonal projection onto $C$. The iteration in Equation (8.1) becomes

$$x^k = P_C\Big(x^{k-1} - \gamma\nabla f_2(x^{k-1})\Big). \tag{8.15}$$

The sequence $\{x^k\}$ converges to a minimizer of $f_2$ over $x \in C$, whenever such minimizers exist.

#### The $CQ$ Algorithm

Let $A$ be a real $I$ by $J$ matrix, $C \subseteq R^J$, and $Q \subseteq R^I$, both closed convex sets. The *split feasibility problem* (SFP) is to find $x$ in $C$ such that $Ax$ is in $Q$. The function

$$f_2(x) = \frac{1}{2}\|P_Q Ax - Ax\|_2^2 \tag{8.16}$$

is convex, differentiable and $\nabla f_2$ is $L$-Lipschitz for $L = \rho(A^T A)$, the spectral radius of $A^T A$. The gradient of $f_2$ is

$$\nabla f_2(x) = A^T(I - P_Q)Ax. \tag{8.17}$$

We want to minimize the function $f_2(x)$ over $x$ in $C$, or, equivalently, to minimize the function $f(x) = \iota_C(x)+f_2(x)$. The projected gradient descent algorithm has the iterative step

$$x^k = P_C\Big(x^{k-1} - \gamma A^T(I - P_Q)Ax^{k-1}\Big); \tag{8.18}$$

this iterative method was called the $CQ$-algorithm in [41, 42]. The sequence $\{x^k\}$ converges to a solution whenever $f_2$ has a minimum on the set $C$.

**The Projected Landweber Algorithm**

The problem is to minimize the function

$$f_2(x) = \frac{1}{2}\|Ax - b\|_2^2,$$

over $x \in C$. This is a special case of the SFP and we can use the $CQ$-algorithm, with $Q = \{b\}$. The resulting iteration is the projected Landweber algorithm; when $C = R^J$ it becomes the Landweber algorithm.

## 8.7 Minimizing $f_2$ over a Linear Manifold

Suppose that we want to minimize $f_2$ over the set of $x$ in the linear manifold $M = S + p$, where $S$ is a subspace of $R^J$ of dimension $I < J$ and $p$ is a fixed vector. Let $A$ be an $I$ by $J$ matrix such that the $I$ columns of $A^T$ form a basis for $S$. For each $z \in R^I$ let

$$d(z) = f_2(A^T z + p),$$

so that $d$ is convex, differentiable, and its gradient,

$$\nabla d(z) = A\nabla f_2(A^T z + p),$$

is $K = \rho(A^T A)L$-Lipschitz. The iteration

$$z^k = z^{k-1} - \gamma\nabla d(z^{k-1}) \tag{8.19}$$

converges to a minimizer of $d$ over all $z$ in $R^I$, whenever minimizers exist, for any $\gamma$ in the interval $(0, \frac{1}{K})$.

From Equation (8.19) we get

$$x^k = x^{k-1} - \gamma A^T A\nabla f_2(x^{k-1}), \tag{8.20}$$

with $x^k = A^T z^k + p$. The sequence $\{x^k\}$ converges to a minimizer of $f_2$ over all $x$ in $M$.

Suppose now that we begin with an algorithm having the iterative step

$$x^k = x^{k-1} - \gamma A^T A \nabla f_2(x^{k-1}), \tag{8.21}$$

where $A$ is any real $I$ by $J$ matrix having rank $I$. Let $x^0$ be in the range of $A^T$, so that $x^0 = A^T z^0$, for some $z^0 \in R^I$. Then each $x^k = A^T z^k$ is again in the range of $A^T$, and we have

$$A^T z^k = A^T z^{k-1} - \gamma A^T A \nabla f_2(A^T z^{k-1}). \tag{8.22}$$

With $d(z) = f_2(A^T z)$, we can write Equation (8.22) as

$$A^T \Big( z^k - (z^{k-1} - \gamma \nabla d(z^{k-1})) \Big) = 0. \tag{8.23}$$

Since $A$ has rank $I$, $A^T$ is one-to-one, so that

$$z^k - z^{k-1} - \gamma \nabla d(z^{k-1}). \tag{8.24}$$

The sequence $\{z^k\}$ converges to a minimizer of $d$, over all $z \in R^I$, whenever such minimizers exist, for $0 < \gamma < \frac{1}{K}$. Therefore, the sequence $\{x^k\}$ converges to a minimizer of $f_2$ over all $x$ in the range of $A^T$.

## 8.8 Feasible-Point Algorithms

Suppose that we want to minimize a convex differentiable function $f(x)$ over $x$ such that $Ax = b$, where $A$ is an $I$ by $J$ full-rank matrix, with $I < J$. If $Ax^k = b$ for each of the vectors $\{x^k\}$ generated by the iterative algorithm, we say that the algorithm is a *feasible-point* method.

### 8.8.1 The Projected Gradient Algorithm

Let $C$ be the feasible set of all $x$ in $R^J$ such that $Ax = b$. For every $z$ in $R^J$, we have

$$P_C z = P_{NS(A)} z + A^T (AA^T)^{-1} b, \tag{8.25}$$

where $NS(A)$ is the null space of $A$. Using

$$P_{NS(A)} z = z - A^T (AA^T)^{-1} Az, \tag{8.26}$$

we have

$$P_C z = z + A^T (AA^T)^{-1} (b - Az). \tag{8.27}$$

For the *projected gradient algorithm* the iteration in Equation (8.1) becomes

$$x^k = x^{k-1} - \gamma P_{NS(A)} \nabla f(x^{k-1}), \tag{8.28}$$

which converges to a solution for any $\gamma$ in $(0, \frac{1}{L})$, whenever solutions exist.

In the next subsection we present a somewhat simpler approach.

### 8.8.2 The Reduced Gradient Algorithm

Let $x^0$ be a *feasible point*, that is, $Ax^0 = b$. Then $x = x^0 + p$ is also feasible if $p$ is in the null space of $A$, that is, $Ap = 0$. Let $Z$ be a $J$ by $J - I$ 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

$$\phi(v) = f(x^0 + Zv)$$

is minimized. We can apply to the function $\phi(v)$ the steepest descent method, or the Newton-Raphson method, or any other minimization technique.

The steepest descent method, applied to $\phi(v)$, is called the *reduced steepest descent algorithm* [129]. The gradient of $\phi(v)$, also called the *reduced gradient*, is

$$\nabla\phi(v) = Z^T \nabla f(x),$$

where $x = x^0 + Zv$; the gradient operator $\nabla\phi$ is then $K$-Lipschitz, for $K = \rho(A^T A)L$.

Let $x^0$ be feasible. The iteration in Equation (8.1) now becomes

$$v^k = v^{k-1} - \gamma \nabla\phi(v^{k-1}), \tag{8.29}$$

so that the iteration for $x^k = x^0 + Zv^k$ is

$$x^k = x^{k-1} - \gamma Z Z^T \nabla f(x^{k-1}). \tag{8.30}$$

The vectors $x^k$ are feasible and the sequence $\{x^k\}$ converges to a solution, whenever solutions exist, for any $0 < \gamma < \frac{1}{K}$.

### 8.8.3 The Reduced Newton-Raphson Method

The same idea can be applied to the Newton-Raphson method. The Newton-Raphson method, applied to $\phi(v)$, is called the *reduced Newton-Raphson method* [129]. The Hessian matrix of $\phi(v)$, also called the *reduced Hessian matrix*, is

$$\nabla^2\phi(v) = Z^T \nabla^2 f(c) Z,$$

so that the reduced Newton-Raphson iteration becomes

$$x^k = x^{k-1} - Z\Big(Z^T \nabla^2 f(x^{k-1}) Z\Big)^{-1} Z^T \nabla f(x^{k-1}). \tag{8.31}$$

Let $c^0$ be feasible. Then each $x^k$ is feasible. The sequence $\{x^k\}$ is not guaranteed to converge.

# Chapter 9

# Alternating Minimization and SUMMA

## 9.1 The Context

As we have seen, both the EMML and the SMART are best derived as alternating minimization (AM) algorithms. The idea of using the AM framework for EMML is due to Vardi, Shepp and Kaufman [149]. The main reference for alternating minimization is the paper [70] of Csiszár and Tusnády. As the authors of [149] remark, the geometric argument in [70] is "deep, though hard to follow". Over the years, I have returned to [70] several times, hoping to simplify that paper and get a better understanding of what they are saying. Finally, in 2011, I managed to clear away the clutter and get to the basics of that paper. This chapter is the result. Once again, it counts as a private victory; I have to assume others have performed the same clearing out and what I have here will never be published. As we shall see, all AM methods for which the five-point property of [70] holds fall into the SUMMA class. Consequently, both the SMART and EMML algorithms are also SUMMA algorithms.

## 9.2 Alternating Minimization

The alternating minimization (AM) iteration of Csiszár and Tusnády [70] provides a useful framework for the derivation of iterative optimization algorithms. In this section we discuss their five-point property and use it to obtain a somewhat simpler proof of convergence for their AM algorithm.

### 9.2.1 The AM Framework

Suppose that $P$ and $Q$ are arbitrary non-empty sets and the function $\Theta(p,q)$ satisfies $-\infty < \Theta(p,q) \leq +\infty$, for each $p \in P$ and $q \in Q$. We assume that, for each $p \in P$, there is $q \in Q$ with $\Theta(p,q) < +\infty$. Therefore, $b = \inf_{p\in P,\, q\in Q} \Theta(p,q) < +\infty$. We assume also that $b > -\infty$; in many applications, the function $\Theta(p,q)$ is non-negative, so this additional assumption is unnecessary. We do not always assume there are $\hat{p} \in P$ and $\hat{q} \in Q$ such that $\Theta(\hat{p},\hat{q}) = b$; when we do assume that such a $\hat{p}$ and $\hat{q}$ exist, we will not assume that $\hat{p}$ and $\hat{q}$ are unique with that property. The objective is to generate a sequence $\{(p^n, q^n)\}$ such that $\Theta(p^n, q^n) \to b$.

### 9.2.2 The AM Iteration

The general AM method proceeds in two steps: we begin with some $q^0$, and, having found $q^n$, we

- **1.** minimize $\Theta(p, q^n)$ over $p \in P$ to get $p = p^{n+1}$, and then
- **2.** minimize $\Theta(p^{n+1}, q)$ over $q \in Q$ to get $q = q^{n+1}$.

In certain applications we consider the special case of alternating cross-entropy minimization. In that case, the vectors $p$ and $q$ are non-negative, and the function $\Theta(p,q)$ will have the value $+\infty$ whenever there is an index $j$ such that $p_j > 0$, but $q_j = 0$. It is important for those particular applications that we select $q^0$ with all positive entries. We therefore assume, for the general case, that we have selected $q^0$ so that $\Theta(p, q^0)$ is finite for all $p$.

The sequence $\{\Theta(p^n, q^n)\}$ is decreasing and bounded below by $b$, since we have

$$\Theta(p^n, q^n) \geq \Theta(p^{n+1}, q^n) \geq \Theta(p^{n+1}, q^{n+1}). \tag{9.1}$$

Therefore, the sequence $\{\Theta(p^n, q^n)\}$ converges to some $B \geq b$. Without additional assumptions, we can say little more.

We know two things:

$$\Theta(p^{n+1}, q^n) - \Theta(p^{n+1}, q^{n+1}) \geq 0, \tag{9.2}$$

and

$$\Theta(p^n, q^n) - \Theta(p^{n+1}, q^n) \geq 0. \tag{9.3}$$

Equation 9.3 can be strengthened to

$$\Theta(p, q^n) - \Theta(p^{n+1}, q^n) \geq 0. \tag{9.4}$$

We need to make these inequalities more precise.

### 9.2.3 The Five-Point Property for AM

The five-point property is the following: for all $p \in P$ and $q \in Q$ and $n = 1, 2, ...$

**The Five-Point Property**

$$\Theta(p, q) + \Theta(p, q^{n-1}) \geq \Theta(p, q^n) + \Theta(p^n, q^{n-1}). \tag{9.5}$$

### 9.2.4 The Main Theorem for AM

We want to find sufficient conditions for the sequence $\{\Theta(p^n, q^n)\}$ to converge to $b$, that is, for $B = b$. The following is the main result of [70].

**Theorem 9.1** *If the five-point property holds then $B = b$.*

**Proof:** Suppose that $B > b$. Then there are $p'$ and $q'$ such that $B > \Theta(p', q') \geq b$. From the five-point property we have

$$\Theta(p', q^{n-1}) - \Theta(p^n, q^{n-1}) \geq \Theta(p', q^n) - \Theta(p', q'), \tag{9.6}$$

so that

$$\Theta(p', q^{n-1}) - \Theta(p', q^n) \geq \Theta(p^n, q^{n-1}) - \Theta(p', q') \geq 0. \tag{9.7}$$

All the terms being subtracted can be shown to be finite. It follows that the sequence $\{\Theta(p', q^{n-1})\}$ is decreasing, bounded below, and therefore convergent. The right side of Equation (9.7) must therefore converge to zero, which is a contradiction. We conclude that $B = b$ whenever the five-point property holds in AM. ∎

### 9.2.5 The Three- and Four-Point Properties

In [70] the five-point property is related to two other properties, the three- and four-point properties. This is a bit peculiar for two reasons: first, as we have just seen, the five-point property is sufficient to prove the main theorem; and second, these other properties involve a second function, $\Delta : P \times P \to [0, +\infty]$, with $\Delta(p, p) = 0$ for all $p \in P$. The three- and four-point properties jointly imply the five-point property, but to get the converse, we need to use the five-point property to define this second function; it can be done, however.

The three-point property is the following:

**The Three-Point Property**

$$\Theta(p, q^n) - \Theta(p^{n+1}, q^n) \geq \Delta(p, p^{n+1}), \tag{9.8}$$

for all $p$. The four-point property is the following:

**The Four-Point Property**

$$\Delta(p, p^{n+1}) + \Theta(p, q) \geq \Theta(p, q^{n+1}), \tag{9.9}$$

for all $p$ and $q$.

It is clear that the three- and four-point properties together imply the five-point property. We show now that the three-point property and the four-point property are implied by the five-point property. For that purpose we need to define a suitable $\Delta(p, \tilde{p})$. For any $p$ and $\tilde{p}$ in $P$ define

$$\Delta(p, \tilde{p}) = \Theta(p, q(\tilde{p})) - \Theta(p, q(p)), \tag{9.10}$$

where $q(p)$ denotes a member of $Q$ satisfying $\Theta(p, q(p)) \leq \Theta(p, q)$, for all $q$ in $Q$. Clearly, $\Delta(p, \tilde{p}) \geq 0$ and $\Delta(p, p) = 0$. The four-point property holds automatically from this definition, while the three-point property follows from the five-point property. Therefore, it is sufficient to discuss only the five-point property when speaking of the AM method.

In the next two sections we discuss the SMART and EMML algorithms, two important instances of alternating minimization.

## 9.3 The SMART

In this section we consider the *simultaneous multiplicative algebraic reconstruction technique* (SMART) as an example of AM.

### 9.3.1 The Kullback-Leibler Distance

The Kullback-Leibler distance plays a fundamental role in the development of both the SMART and the EMML algorithms.

For $\alpha > 0$ and $\beta > 0$, the Kullback-Leibler distance, $KL(\alpha, \beta)$, is defined as

$$KL(\alpha, \beta) = \alpha \log \frac{\alpha}{\beta} + \beta - \alpha. \tag{9.11}$$

In addition, $KL(0, 0) = 0$, $KL(\alpha, 0) = +\infty$ and $KL(0, \beta) = \beta$. The KL distance is then extended to non-negative vectors coordinate-wise.

One of the most useful facts about the KL distance is contained in the following lemma.

**Lemma 9.1** *For non-negative vectors $x$ and $z$, with $z_+ = \sum_{j=1}^{J} z_j > 0$, we have*

$$KL(x, z) = KL(x_+, z_+) + KL(x, \frac{x_+}{z_+} z). \tag{9.12}$$

### 9.3.2 Background

What is usually called the simultaneous multiplicative algebraic reconstruction technique (SMART) was discovered in 1972, independently, by Darroch and Ratcliff [71], working in statistics, and by Schmidlin [138] in medical imaging. The SMART provides another example of alternating minimization having the three- and four-point properties.

Darroch and Ratcliff called their algorithm *generalized iterative scaling*. It was designed to calculate the entropic projection of one probability vector onto a family of probability vectors with a pre-determined marginal distribution. They did not consider the more general problems of finding a non-negative solution of a non-negative system of linear equations $y = Px$, or of minimizing a function; they did not, therefore, consider what happens in the inconsistent case, in which the system of equations $y = Px$ has no non-negative solutions. This issue was resolved in [30], where it was shown that the SMART minimizes the function $f(x) = KL(Px, y)$, over non-negative vectors $x$. Here $y$ is a vector with positive entries, and $P$ is a matrix with non-negative entries, such that $s_j = \sum_{i=1}^{I} P_{ij} > 0$ for all $j$. This function is continuous in the variable $x$ and has bounded level sets, so there is at least one minimizer; call it $\hat{x}$. The vector $P\hat{x}$ is unique, even if the vector $\hat{x}$ is not. For notational convenience we shall assume that $s_j = 1$ for all $j$. If this is not the case initially, we replace $P_{ij}$ with $P_{ij}/s_j$ and $x_j$ with $x_j s_j$; the product $Px$ is unchanged.

### 9.3.3 Some Notation for SMART

Let $\mathcal{X}$ be the set of all $x \geq 0$ for which the vector $Px$ has only positive entries. For each $x \in \mathcal{X}$, let $t(x)$ and $r(x)$ be the $I$ by $J$ arrays with entries

$$t(x)_{ij} = x_j P_{ij}, \tag{9.13}$$

and

$$r(x)_{ij} = x_j P_{ij} y_i / (Px)_i. \tag{9.14}$$

We then let

$$\mathcal{R} = \{r = \{r_{ij} \geq 0\} | \sum_{j=1}^{J} r_{ij} = y_i, \text{ for } i = 1, 2, ..., I\}, \tag{9.15}$$

and

$$\mathcal{T} = \{t = t(x) | x \in \mathcal{X}\}. \tag{9.16}$$

The sets $\mathcal{R}$ and $\mathcal{T}$ are convex in the space $R^{I+J}$.

### 9.3.4 Pythagorean Identities

Using the following Pythagorean identities we can prove convergence of the SMART algorithm [30, 32]:

$$KL(t(x), r(z)) = KL(t(x), r(x)) + KL(x, z) - KL(Px, Pz); \tag{9.17}$$

and

$$KL(t(x), r(z)) = KL(t(z^*), r(z)) + KL(x, z^*), \tag{9.18}$$

where $x$ and $z$ are arbitrary members of $\mathcal{X}$ and

$$z_j^* = z_j \exp\Big(\sum_{i=1}^{I} P_{ij} \log\Big(\frac{y_i}{(Pz)_i}\Big)\Big), \tag{9.19}$$

for each $j$. Note that

$$KL(Px, y) = KL(t(x), r(x)), \tag{9.20}$$

and

$$KL(x, z) - KL(Px, Pz) \geq 0. \tag{9.21}$$

### 9.3.5 The SMART Iteration

The iterative step of the SMART is to minimize the function $KL(t(x), r(x^{n-1}))$ to get $x = x^n$. The SMART iteration begins with a positive vector $x^0$. Having found the vector $x^{n-1}$, the next vector in the SMART sequence is $x^n = (x^{n-1})^*$, with entries given by

$$x_j^n = (x^{n-1})_j^* = x_j^{n-1} \exp\Big(\sum_{i=1}^{I} P_{ij} \log\Big(\frac{y_i}{(Px^{n-1})_i}\Big)\Big). \tag{9.22}$$

The sequence $\{x^n\}$ converges to the non-negative minimizer of the function $KL(Px, y)$ for which $KL(x, x^0)$ is minimized.

### 9.3.6 The SMART as AM

To put the SMART algorithm into the framework of alternating minimization, we take the sets $Q = \mathcal{R}$ and $P = \mathcal{T}$ as above and let $p^n = t(x^n)$, and $q^n = r(x^n)$. Generic vectors are $p = t(x)$ for some $x \in \mathcal{X}$ and $q = r(z)$ for some $z \in \mathcal{X}$. Then we set

$$\Theta(p, q) = KL(t(x), r(z)), \tag{9.23}$$

and, for arbitrary $p = t(x)$ and $\tilde{p} = t(\tilde{x})$,

$$\Delta(p, \tilde{p}) = KL(t(x), t(\tilde{x})) = KL(x, \tilde{x}). \tag{9.24}$$

From the Pythagorean identity (9.18) we have

$$KL(t(x), r(x^{n-1})) = KL(t(x^n), r(x^{n-1})) + KL(x, x^n) \tag{9.25}$$

so that

$$\Theta(p, q^{n-1}) = \Theta(p^n, q^{n-1}) + \Delta(p, p^n), \tag{9.26}$$

which is then the three-point property. From

$$KL(t(x), r(x^n)) - KL(t(x), r(x)) = KL(x, x^n) - KL(Px, Px^n) \leq KL(x, x^n) \tag{9.27}$$

we have

$$\Delta(p, p^n) \geq \Theta(p, q^n) - \Theta(p, q(p)) \geq \Theta(p, q^n) - \Theta(p, q), \tag{9.28}$$

which is the four-point property.

The iterative step of the SMART is then to minimize the function

$$\Theta(p, q^{n-1}) = KL(t(x), r(x^{n-1})) \tag{9.29}$$

to get $x = x^n = (x^{n-1})^*$. Since the SMART is a particular case of AM for which the five-point property holds, we know that

$$\{KL(Px^n, y)\} \to \inf\{KL(Px, y) \,|x \geq 0\}. \tag{9.30}$$

Using the Pythagorean identities we can show more: the sequence $\{x^n\}$ converges to the non-negative minimizer of the function $KL(Px, y)$ for which $KL(x, x^0)$ is minimized ([30, 32]) .

### 9.3.7 Related work of Csiszár

In [69] Csiszár shows that the generalized iterative scaling method of Darroch and Ratcliff can be formulated in terms of successive entropic projection onto the sets $\mathcal{R}$ and $\mathcal{T}$. In other words, he views their method as an alternating projection method, not as alternating minimization. He derives the generalized iterative scaling algorithm in two steps:

- 1. minimize $KL(r(x), t(x^n))$ to get $r(x^n)$; and then
- 2. minimize $KL(t(x), r(x^n))$ to get $t(x^{n+1})$.

Although [69] appeared five years after [70], Csiszár does not reference [70], nor does he mention alternating minimization, instead basing his convergence proof here on his earlier paper [68], which deals with entropic projection. He is able to make this work because the order of the $t(x^n)$ and $r(x)$ does not matter in the first step. Therefore, the generalized iterative scaling, and, more generally, the SMART, is also an alternating projection algorithm, as well.

## 9.4 The EMML Algorithm

The *expectation maximization maximum likelihood* (EMML) method we discuss here is actually a special case of a more general approach to likelihood maximization, usually called the EM algorithm [74]; the book by McLachnan and Krishnan [120] is a good source for the history of this more general algorithm.

### 9.4.1 Background

It was noticed by Rockmore and Macovski [137] that the image reconstruction problems that arise in medical tomography can be formulated as statistical parameter estimation problems. Following up on this idea, Shepp and Vardi [139] suggested the use of the EM algorithm for solving the reconstruction problem in emission tomography. In [113], 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 [139] for the emission case. In [149], Vardi, Shepp and Kaufman repaired the earlier proof, relying on techniques due to Csiszár and Tusnády [70]. In [114] 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 [30, 31].

The EMML, as a statistical parameter estimation technique, was not originally thought to be connected to any system of linear equations. In [30], it was shown that the EMML algorithm minimizes the function $f(x) = KL(y, Px)$, over non-negative vectors $x$. As in the previous section, $y$ is a vector with positive entries, and $P$ is a matrix with non-negative entries, such that $s_j = \sum_{i=1}^{I} P_{ij} = 1$. Consequently, when the non-negative system of linear equations $Px = y$ has a non-negative solution, the EMML converges to such a solution.

Because $KL(y, Px)$ is continuous in the variable $x$ and has bounded level sets, there is at least one non-negative minimizer; call it $\hat{x}$. The vector $P\hat{x}$ is unique, even if $\hat{x}$ is not.

### 9.4.2 Pythagorean Identities

For each $x \in \mathcal{X}$, let $t(x)$ and $r(x)$ be as previously defined. Using the following Pythagorean identities we can prove convergence of the EMML algorithm [30, 32]:

$$KL(r(x), t(z)) = KL(r(z), t(z)) + KL(r(x), r(z)); \tag{9.31}$$

and

$$KL(r(x), t(z)) = KL(r(x), t(x')) + KL(x', z), \tag{9.32}$$

where $x$ and $z$ are arbitrary members of $\mathcal{X}$ and the entries of $x'$ are defined by

$$x'_j = x_j \sum_{i=1}^{I} P_{ij} \frac{y_i}{(Px)_i}, \tag{9.33}$$

for each $j$. Note that $KL(y, Px) = KL(r(x), t(x))$.

### 9.4.3 The EMML as AM

In the EMML algorithm we minimize the function $KL(r(x^n), t(x))$ to get $x = x^{n+1}$. The EMML iteration begins with a positive vector $x^0$. Having found the vector $x^n$, the next vector in the EMML sequence is $x^{n+1} = (x^n)'$, with entries given by

$$x_j^{n+1} = (x^n)'_j = x_j^n \sum_{i=1}^{I} P_{ij} \Big( \frac{y_i}{(Px^n)_i} \Big) = \sum_{i=1}^{I} r(x^n)_{ij}. \tag{9.34}$$

The sequence $\{x^n\}$ converges to a non-negative minimizer of the function $KL(y, Px)$.

We put the EMML algorithm into an AM framework using $P = \mathcal{R}$, $Q = \mathcal{T}$, $p = r(x)$, $q = t(z)$, $\Theta(p, q) = KL(r(x), t(z))$, and minimizing $KL(r(x), t(x)) = KL(y, Px)$. Using the AM notation, we let $q^{n-1} = t(x^{n-1})$, $p^n = r(x^{n-1})$, $p = r(x)$, $\tilde{p} = r(\tilde{x})$, and $q(p) = t(x')$. At the $n$th step of the EMML algorithm we obtain $p^n = r(x^{n-1})$ by minimizing

$$\Theta(p, q^{n-1}) = KL(r(x), t(x^{n-1})). \tag{9.35}$$

According to the Pythagorean identities (9.31) and (9.32) and Lemma 9.12, we have $x^n = (x^{n-1})'$ and

$$\Theta(p, q^{n-1}) - \Theta(p^n, q^{n-1}) = KL(r(x), r(x^{n-1})) \geq KL(x', (x^{n-1})') = KL(x', x^n). \tag{9.36}$$

With $\Delta(p, \tilde{p})$ defined as

$$\Delta(p, \tilde{p}) = KL(r(x), r(\tilde{x})), \tag{9.37}$$

it follows that

$$\Delta(p, p^n) = KL(r(x), r(x^{n-1})), \tag{9.38}$$

so that

$$\Theta(p, q^{n-1}) - \Theta(p^n, q^{n-1}) \geq \Delta(p, p^n), \tag{9.39}$$

which is the three-point property.

We know that

$$KL(r(x), t(x^n)) - KL(r(x), t(x')) = KL(x', x^n) \tag{9.40}$$

and

$$KL(r(x), r(x^{n-1})) \geq KL(x', x^n), \tag{9.41}$$

from which it follows that

$$KL(r(x), r(x^{n-1})) \geq KL(r(x), t(x^n)) - KL(r(x), t(x')); \tag{9.42}$$

this is the four-point property.

## 9.5 Alternating Bregman Distance Minimization

The general problem of minimizing $\Theta(p, q)$ is simply a minimization of a real-valued function of two variables, $p \in P$ and $q \in Q$. In many cases the function $\Theta(p, q)$ is a distance between $p$ and $q$, either $\|p - q\|_2^2$ or $KL(p, q)$. In the case of $\Theta(p, q) = \|p - q\|_2^2$, each step of the alternating minimization algorithm involves an orthogonal projection onto a closed convex set; both projections are with respect to the same Euclidean distance function. In the case of cross-entropy minimization, we first project $q^n$ onto the set $P$ by minimizing the distance $KL(p, q^n)$ over all $p \in P$, and then project $p^{n+1}$ onto the set $Q$ by minimizing the distance function $KL(p^{n+1}, q)$. This suggests the possibility of using alternating minimization with respect to more general distance functions. We shall focus on Bregman distances.

### 9.5.1 Bregman Distances

Let $f : R^N \to R$ be a Bregman function [15, 61, 21], and so $f(x)$ is convex on its domain and differentiable in the interior of its domain. Then, for $x$ in the domain and $z$ in the interior, we define the Bregman distance $D_f(x, z)$ by

$$D_f(x, z) = f(x) - f(z) - \langle \nabla f(z), x - z \rangle. \tag{9.43}$$

For example, the KL distance is a Bregman distance with associated Bregman function

$$f(x) = \sum_{j=1}^{J} x_j \log x_j \; - x_j. \tag{9.44}$$

Suppose now that $f(x)$ is a Bregman function and $P$ and $Q$ are closed convex subsets of the interior of the domain of $f(x)$. Let $p^{n+1}$ minimize $D_f(p, q^n)$ over all $p \in P$. It follows then that

$$\langle \nabla f(p^{n+1}) - \nabla f(q^n), p - p^{n+1} \rangle \geq 0, \tag{9.45}$$

for all $p \in P$. Since

$$D_f(p, q^n) - D_f(p^{n+1}, q^n) = D_f(p, p^{n+1}) + \langle \nabla f(p^{n+1}) - \nabla f(q^n), p - p^{n+1} \rangle, \tag{9.46}$$

it follows that the three-point property holds, with

$$\Theta(p, q) = D_f(p, q), \tag{9.47}$$

and

$$\Delta(p, \hat{p}) = D_f(p, \tilde{p}). \tag{9.48}$$

To get the four-point property we need to restrict $D_f$ somewhat; we assume from now on that $D_f(p, q)$ is jointly convex, that is, it is convex in the combined vector variable $(p, q)$ (see [9]). Now we can invoke a lemma due to Eggermont and LaRiccia [79].

### 9.5.2 The Eggermont-LaRiccia Lemma

**Lemma 9.2** *Suppose that the Bregman distance $D_f(p, q)$ is jointly convex. Then it has the four-point property.*

**Proof:** By joint convexity we have

$$D_f(p, q) - D_f(p^n, q^n) \geq$$

$$\langle \nabla_1 D_f(p^n, q^n), p - p^n \rangle + \langle \nabla_2 D_f(p^n, q^n), q - q^n \rangle,$$

where $\nabla_1$ denotes the gradient with respect to the first vector variable. Since $q^n$ minimizes $D_f(p^n, q)$ over all $q \in Q$, we have

$$\langle \nabla_2 D_f(p^n, q^n), q - q^n \rangle \geq 0,$$

for all $q$. Also,

$$\langle \nabla_1(p^n, q^n), p - p^n \rangle = \langle \nabla f(p^n) - \nabla f(q^n), p - p^n \rangle.$$

It follows that

$$D_f(p, q^n) - D_f(p, p^n) = D_f(p^n, q^n) + \langle \nabla_1(p^n, q^n), p - p^n \rangle$$

$$\leq D_f(p, q) - \langle \nabla_2 D_f(p^n, q^n), q - q^n \rangle \leq D_f(p, q).$$

Therefore, we have

$$D_f(p, p^n) + D_f(p, q) \geq D_f(p, q^n).$$

This is the four-point property. ∎

We now know that the alternating minimization method works for any Bregman distance that is jointly convex. This includes the Euclidean and the KL distances.

## 9.6 Minimizing a Proximity Function

We present now an example of alternating Bregman distance minimization taken from [40]. The problem is the *convex feasibility problem* (CFP), to find a member of the intersection $C \subseteq R^J$ of finitely many closed convex sets $C_i$, $i = 1, ..., I$, or, failing that, to minimize the proximity function

$$F(x) = \sum_{i=1}^{I} D_i(\overleftarrow{P}_i x, x), \tag{9.49}$$

where $f_i$ are Bregman functions for which $D_i$, the associated Bregman distance, is jointly convex, and $\overleftarrow{P}_i x$ are the *left* Bregman projection of $x$ onto the set $C_i$, that is, $\overleftarrow{P}_i x \in C_i$ and $D_i(\overleftarrow{P}_i x, x) \leq D_i(z, x)$, for all $z \in C_i$. Because each $D_i$ is jointly convex, the function $F(x)$ is convex.

The problem can be formulated as an alternating minimization, where $P \subseteq R^{IJ}$ is the product set $P = C_1 \times C_2 \times ... \times C_I$. A typical member of $P$ has the form $p = (c^1, c^2, ..., c^I)$, where $c^i \in C_i$, and $Q \subseteq R^{IJ}$ is the *diagonal* subset, meaning that the elements of $Q$ are the $I$-fold product of a single $x$; that is $Q = \{d(x) = (x, x, ..., x) \in R^{IJ}\}$. We then take

$$\Theta(p, q) = \sum_{i=1}^{I} D_i(c^i, x), \tag{9.50}$$

and $\Delta(p, \tilde{p}) = \Theta(p, \tilde{p})$.

In [56] a similar iterative algorithm was developed for solving the CFP, using the same sets $P$ and $Q$, but using alternating projection, rather than alternating minimization. Now it is not necessary that the Bregman distances be jointly convex. Each iteration of their algorithm involves two steps:

- 1. minimize $\sum_{i=1}^{I} D_i(c^i, x^n)$ over $c^i \in C_i$, obtaining $c^i = \overleftarrow{P}_i x^n$, and then
- 2. minimize $\sum_{i=1}^{I} D_i(x, \overleftarrow{P}_i x^n)$.

Because this method is an alternating projection approach, it converges only when the CFP has a solution, whereas the previous alternating minimization method minimizes $F(x)$, even when the CFP has no solution.

### 9.6.1 Right and Left Projections

Because Bregman distances $D_f$ are not generally symmetric, we can speak of *right* and *left* Bregman projections onto a closed convex set. For any allowable vector $x$, the *left* Bregman projection of $x$ onto $C$, if it exists, is the vector $\overleftarrow{P}_C x \in C$ satisfying the inequality $D_f(\overleftarrow{P}_C x, x) \leq D_f(c, x)$, for all $c \in C$. Similarly, the *right* Bregman projection is the vector $\overrightarrow{P}_C x \in C$ satisfying the inequality $D_f(x, \overrightarrow{P}_C x) \leq D_f(x, c)$, for any $c \in C$.

The alternating minimization approach described above to minimize the proximity function

$$F(x) = \sum_{i=1}^{I} D_i(\overleftarrow{P}_i x, x) \tag{9.51}$$

can be viewed as an alternating projection method, but employing both right and left Bregman projections.

Consider the problem of finding a member of the intersection of two closed convex sets $C$ and $D$. We could proceed as follows: having found $x^n$, minimize $D_f(x^n, d)$ over all $d \in D$, obtaining $d = \overrightarrow{P}_D x^n$, and then minimize $D_f(c, \overrightarrow{P}_D x^n)$ over all $c \in C$, obtaining $c = x^{n+1} = \overleftarrow{P}_C \overrightarrow{P}_D x^n$. The objective of this algorithm is to minimize $D_f(c, d)$ over all $c \in C$ and $d \in D$; such a minimum may not exist, of course.

In [10] the authors note that the alternating minimization algorithm of [40] involves right and left Bregman projections, which suggests to them iterative methods involving a wider class of operators that they call "Bregman retractions".

## 9.7 More Proximity Function Minimization

Proximity function minimization and right and left Bregman projections play a role in a variety of iterative algorithms. We survey several of them in this section.

### 9.7.1 Cimmino's Algorithm

Our objective here is to find an exact or approximate solution of the system of $I$ linear equations in $J$ unknowns, written $Ax = b$. For each $i$ let

$$C_i = \{z | (Az)_i = b_i\}, \tag{9.52}$$

and $P_i x$ be the orthogonal projection of $x$ onto $C_i$. Then

$$(P_i x)_j = x_j + \alpha_i A_{ij}(b_i - (Ax)_i), \tag{9.53}$$

where

$$(\alpha_i)^{-1} = \sum_{j=1}^{J} A_{ij}^2. \tag{9.54}$$

Let

$$F(x) = \sum_{i=1}^{I} \|P_i x - x\|_2^2. \tag{9.55}$$

Using alternating minimization on this proximity function gives Cimmino's algorithm, with the iterative step

$$x_j^{n+1} = x_j^n + \frac{1}{I} \sum_{i=1}^{I} \alpha_i A_{ij} (b_i - (Ax^n)_i). \tag{9.56}$$

### 9.7.2 Simultaneous Projection for Convex Feasibility

Now we let $C_i$ be any closed convex subsets of $R^J$ and define $F(x)$ as in the previous section. Again, we apply alternating minimization. The iterative step of the resulting algorithm is

$$x^{n+1} = \frac{1}{I} \sum_{i=1}^{I} P_i x^n. \tag{9.57}$$

The objective here is to minimize $F(x)$, if there is a minimum.

### 9.7.3 The EMML Revisited

As in our earlier discussion of the SMART and EMML methods, we want an exact or approximate solution of the system $y = Px$. For each $i$, let

$$C_i = \{z \geq 0 \,|\, (Pz)_i = y_i\}. \tag{9.58}$$

The left entropic projection of $x > 0$ onto $C_i$ is the vector that minimizes $KL(c_i, x)$, over all $c_i \in C_i$; unfortunately, we typically cannot calculate this projection in closed form. Instead, we define the distances

$$D_i(z, x) = \sum_{j=1}^{J} P_{ij} KL(z_j, x_j), \tag{9.59}$$

and calculate the associated left projections $\overleftarrow{P}_i x$ onto the sets $C_i$. We then have $D_i(\overleftarrow{P}_i x, x) \leq D_i(c_i, x)$, for all $c_i \in C_i$, with $\overleftarrow{P}_i x$ given in closed form by

$$(\overleftarrow{P}_i x)_j = x_j \frac{y_i}{(Px)_i}, \tag{9.60}$$

for each $j$. Note that, for the distances $D_i$ and these sets $C_i$, the left and right projections are the same; that is $\overleftarrow{P}_i x = \overrightarrow{P}_i x$. Applying alternating minimization to the proximity function

$$F(x) = \sum_{i=1}^{I} \sum_{j=1}^{J} P_{ij} KL(\overleftarrow{P}_i x, x), \tag{9.61}$$

we obtain the iterative step

$$x_j^{n+1} = x_j^n \sum_{i=1}^{I} P_{ij} \frac{y_i}{(Px^n)_i}, \tag{9.62}$$

which is the EMML iteration.

### 9.7.4 The SMART

Now we define the proximity function $F(x)$ to be

$$F(x) = \sum_{i=1}^{I} \sum_{j=1}^{J} P_{ij} KL(x, \overrightarrow{P}_i x). \tag{9.63}$$

Applying alternating minimization and using the fact that $\overleftarrow{P}_i x = \overrightarrow{P}_i x$, we discover that the resulting iterative step is that of the SMART.

### 9.7.5 The Bauschke-Combettes-Noll Problem

In [11] Bauschke, Combettes and Noll consider the following problem: minimize the function

$$\Theta(p, q) = \Lambda(p, q) = \phi(p) + \psi(q) + D_f(p, q), \tag{9.64}$$

where $\phi$ and $\psi$ are convex on $R^J$, $D = D_f$ is a Bregman distance, and $P = Q$ is the interior of the domain of $f$. They assume that

$$b = \inf_{(p,q)} \Lambda(p, q) > -\infty, \tag{9.65}$$

and seek a sequence $\{(p^n, q^n)\}$ such that $\{\Lambda(p^n, q^n)\}$ converges to $b$. The sequence is obtained by the AM method, as in our previous discussion. They prove that, if the Bregman distance is jointly convex, then $\{\Lambda(p^n, q^n)\} \downarrow b$. In this subsection we obtain this result by showing that $\Lambda(p, q)$ has the five-point property whenever $D = D_f$ is jointly convex. Our proof is loosely based on the proof of the Eggermont-LaRiccia lemma.

The five-point property for $\Lambda(p, q)$ is

$$\Lambda(p, q^{n-1}) - \Lambda(p^n, q^{n-1}) \geq \Lambda(p, q^n) - \Lambda(p, q). \tag{9.66}$$

A simple calculation shows that the inequality in (9.66) is equivalent to

$$\Lambda(p,q) - \Lambda(p^n,q^n) \geq D(p,q^n) + D(p^n,q^{n-1}) - D(p,q^{n-1}) - D(p^n,q^n) \quad (9.67)$$

By the joint convexity of $D(p,q)$ and the convexity of $\phi$ and $\psi$ we have

$$\Lambda(p,q) - \Lambda(p^n,q^n) \geq \langle \nabla_p \Lambda(p^n,q^n), p - p^n \rangle + \langle \nabla_q \Lambda(p^n,q^n), q - q^n \rangle, \quad (9.68)$$

where $\nabla_p \Lambda(p^n,q^n)$ denotes the gradient of $\Lambda(p,q)$, with respect to $p$, evaluated at $(p^n,q^n)$.

Since $q^n$ minimizes $\Lambda(p^n,q)$, it follows that

$$\langle \nabla_q \Lambda(p^n,q^n), q - q^n \rangle = 0, \quad (9.69)$$

for all $q$. Therefore,

$$\Lambda(p,q) - \Lambda(p^n,q^n) \geq \langle \nabla_p \Lambda(p^n,q^n), p - p^n \rangle . \quad (9.70)$$

We have

$$\langle \nabla_p \Lambda(p^n,q^n), p - p^n \rangle = \langle \nabla f(p^n) - \nabla f(q^n), p - p^n \rangle + \langle \nabla \phi(p^n), p - p^n \rangle \quad (9.71)$$

Since $p^n$ minimizes $\Lambda(p,q^{n-1})$, we have

$$\nabla_p \Lambda(p^n,q^{n-1}) = 0, \quad (9.72)$$

or

$$\nabla \phi(p^n) = \nabla f(q^{n-1}) - \nabla f(p^n), \quad (9.73)$$

so that

$$\langle \nabla_p \Lambda(p^n,q^n), p - p^n \rangle = \langle \nabla f(q^{n-1}) - \nabla f(q^n), p - p^n \rangle \quad (9.74)$$

$$= D(p,q^n) + D(p^n,q^{n-1}) - D(p,q^{n-1}) - D(p^n,q^n). \quad (9.75)$$

Using (9.70) we obtain the inequality in (9.67). This shows that $\Lambda(p,q)$ has the five-point property whenever the Bregman distance $D = D_f$ is jointly convex.

From our previous discussion of AM, we conclude that the sequence $\{\Lambda(p^n,q^n)\}$ converges to $b$; this is Corollary 4.3 of [11].

In [54] it was shown that, in certain cases, the expectation maximization maximum likelihood (EM) method involves alternating minimization of a function of the form $\Lambda(p,q)$.

## 9.8 The SUMMA

We turn now to an apparently unrelated problem, to minimize a function $f : S \to (-\infty, \infty]$ over a (not necessarily proper) subset $C$ of $S$. At the $n$th step of a *sequential unconstrained minimization* method, we obtain $x^n$ by minimizing the function

$$G_n(x) = f(x) + g_n(x), \tag{9.76}$$

where the auxiliary function $g_n(x)$ is appropriately chosen. If $C$ is a proper subset of $S$ we may force $g_n(x) = +\infty$ for $x$ not in $C$, as in the barrier-function methods; then each $x^n$ will lie in $C$.

The objective is to select the $g_n(x)$ so that the sequence $\{x^n\}$ converges to a solution of the problem, or failing that, at least to have the sequence $\{f(x^n)\}$ converging to the infimum of $f(x)$ over $x$ in $C$.

In [46] we presented a particular class of sequential unconstrained minimization methods called SUMMA. As we showed in that paper, this class is broad enough to contain barrier-function methods, proximal minimization methods, and the simultaneous multiplicative algebraic reconstruction technique (SMART). By reformulating the problem, the penalty-function methods can also be shown to be members of the SUMMA class. When [46] was written, we were not able to include the *expectation maximization maximum likelihood* (EMML) method [139] within the SUMMA class. As we shall see shortly, any AM problem with the five-point property can be reformulated as a SUMMA problem; therefore the EMML, which is such an AM algorithm, must also be a SUMMA algorithm.

For a method to be in the SUMMA class we require that $x^n \in C$ for each $n$ and that each auxiliary function $g_n(x)$ satisfy the inequalities

$$0 \leq g_{n+1}(x) \leq G_n(x) - G_n(x^n), \tag{9.77}$$

for all $x$. Note that it follows that $g_{n+1}(x^n) = 0$, for all $n$. We assume, throughout this section, that the inequality in (9.77) holds for each $n$. We also assume that $\inf_{x \in C} f(x) = b > -\infty$. The next two results are taken from [46].

**Proposition 9.1** *The sequence $\{f(x^n)\}$ is non-increasing and the sequence $\{g_n(x^n)\}$ converges to zero.*

**Proof:** We have

$$f(x^{n+1}) + g_{n+1}(x^{n+1}) = G_{n+1}(x^{n+1}) \leq G_{n+1}(x^n) = f(x^n). \tag{9.78}$$

∎

**Theorem 9.2** *The sequence $\{f(x^n)\}$ converges to $b$.*

**Proof:** Suppose that there is $\delta > 0$ such that $f(x^n) \geq b + 2\delta$, for all $n$. Then there is $z \in C$ such that $f(x^n) \geq f(z) + \delta$, for all $n$. From the inequality in (9.77) we have

$$g_n(z) - g_{n+1}(z) \geq f(x^n) + g_n(x^n) - f(z) \geq f(x^n) - f(z) \geq \delta, \quad (9.79)$$

for all $n$. But this cannot happen; the successive differences of a non-increasing sequence of non-negative terms must converge to zero. ∎

## 9.9 Examples of SUMMA

In this section we present several examples of SUMMA.

### 9.9.1 Barrier-Function Methods

Let $b(x) : R^J \to (-\infty, +\infty]$ be continuous, with effective domain the set

$$D = \{x | \, b(x) < +\infty\}.$$

The goal is to minimize the objective function $f(x)$, over $x$ in the closed set $C = \overline{D}$, the closure of $D$. In the barrier-function method, we minimize

$$f(x) + \frac{1}{n} b(x) \quad (9.1)$$

over $x$ in $D$ to get $x^n$. Each $x^n$ lies within $D$, so the method is an interior-point algorithm. If the sequence $\{x^n\}$ converges, the limit vector $x^*$ will be in $C$ and $f(x^*) = f(\hat{x})$.

The iterative step of the barrier-function method can be formulated as follows: minimize

$$f(x) + [(n-1)f(x) + b(x)] \quad (9.2)$$

to get $x^n$. Since, for $n = 2, 3, ...$, the function

$$(n-1)f(x) + b(x) \quad (9.3)$$

is minimized by $x^{n-1}$, the function

$$g_n(x) = (n-1)f(x) + b(x) - (n-1)f(x^{n-1}) - b(x^{n-1}) \quad (9.4)$$

is non-negative, and $x^n$ minimizes the function

$$G_n(x) = f(x) + g_n(x). \quad (9.5)$$

From

$$G_n(x) = f(x) + (n-1)f(x) + b(x) - (n-1)f(x^{n-1}) - b(x^{n-1}), \quad (9.6)$$

it follows that

$$G_n(x) - G_n(x^n) = nf(x) + b(x) - nf(x^n) - b(x^n) = g_{n+1}(x), \qquad (9.7)$$

so that $g_{n+1}(x)$ satisfies the condition in (9.77). This shows that the barrier-function method is a particular case of SUMMA.

### 9.9.2 Penalty-Function Methods

Once again, we want to minimize $f(x)$ over $x \in C$. In penalty-function methods the $n$th step is to minimize

$$f(x) + np(x), \qquad (9.8)$$

where $p(x) > 0$ for $x$ not in $C$ and $p(x) = 0$ for $x \in C$. To show that penalty-function methods can be viewed as members of the SUMMA class, we reformulate these methods as barrier-function methods. In order to relate penalty-function methods to barrier-function methods, we note that minimizing $f(x) + np(x)$ is equivalent to minimizing $p(x) + \frac{1}{n}f(x)$. This is the form of the barrier-function iteration, with $p(x)$ now in the role previously played by $f(x)$, and $f(x)$ now in the role previously played by $b(x)$. We are not concerned here with the effective domain of $f(x)$.

### 9.9.3 Proximity-Function Minimization

Let $f : R^J \to (-\infty, +\infty]$ be closed, proper, convex and differentiable. Let $h$ be a closed proper convex function, with effective domain $D$, that is differentiable on the nonempty open convex set int $D$. Assume that $f(x)$ is finite on $C = \overline{D}$ and attains its minimum value on $C$ at $\hat{x}$. The corresponding *Bregman distance* $D_h(x, z)$ is defined for $x$ in $D$ and $z$ in int $D$ by

$$D_h(x, z) = h(x) - h(z) - \langle \nabla h(z), x - z \rangle. \qquad (9.9)$$

Note that $D_h(x, z) \geq 0$ always. If $h$ is essentially strictly convex, then $D_h(x, z) = 0$ implies that $x = z$. Our objective is to minimize $f(x)$ over $x$ in $C = \overline{D}$.

At the $n$th step of the *proximal minimization algorithm* (PMA) [38, 61], we minimize the function

$$G_n(x) = f(x) + D_h(x, x^{n-1}), \qquad (9.10)$$

to get $x^n$. The function

$$g_n(x) = D_h(x, x^{n-1}) \qquad (9.11)$$

is non-negative and $g_n(x^{n-1}) = 0$. We assume that each $x^n$ lies in int $D$.

The PMA is a particular case of the SUMMA. We remind the reader that $f(x)$ is now assumed to be convex and differentiable, so that the Bregman distance $D_f(x,z)$ is defined and non-negative, for all $x$ in $D$ and $z$ in int$D$.

**Lemma 9.1** *For each $n$ we have*

$$G_n(x) = G_n(x^n) + D_f(x, x^n) + D_h(x, x^n). \tag{9.12}$$

**Proof:** Since $x^n$ minimizes $G_n(x)$ within the set $D$, we have

$$0 = \nabla f(x^n) + \nabla h(x^n) - \nabla h(x^{n-1}). \tag{9.13}$$

Then

$$G_n(x) - G_n(x^n) = f(x) - f(x^n) + h(x) - h(x^n) - \langle \nabla h(x^{n-1}), x - x^n \rangle \tag{9.14}$$

Now substitute, using Equation (9.13) and the definition of Bregman distances. ∎

It follows from Lemma 9.1 that

$$G_n(x) - G_n(x^n) = g_{n+1}(x) + D_f(x, x^n). \tag{9.15}$$

### 9.9.4 The Simultaneous MART

It follows from the Pythagorean identities established in [30] that the SMART can also be formulated as a particular case of the SUMMA. From the identities established for the SMART in [30], we know that the iterative step of SMART can be expressed as follows: minimize the function

$$G_n(x) = KL(Px, y) + KL(x, x^{n-1}) - KL(Px, Px^{n-1}) \tag{9.16}$$

to get $x^n$. According to Lemma 9.12, the function

$$g_n(x) = KL(x, x^{n-1}) - KL(Px, Px^{n-1}) \tag{9.17}$$

is non-negative, since $s_j = 1$. The $g_n(x)$ are defined for all non-negative $x$; that is, the set $D$ is the closed non-negative orthant in $R^J$. Each $x^n$ is a positive vector.

It was shown in [30] that

$$G_n(x) = G_n(x^n) + KL(x, x^n), \tag{9.18}$$

from which it follows immediately that SMART is a particular case of SUMMA. Consequently, the sequence $\{KL(Px^n, y)\}$ converges to the infimum of the function $KL(Px, y)$ over all $x \in \mathcal{X}$. The infimum is always attained at some $x \geq 0$ in the closure of $\mathcal{X}$ and it can be shown that the sequence $\{x^n\}$ converges to a minimizer of $KL(Px, y)$ over $x$ in the closure of $\mathcal{X}$ ([30, 32]).

## 9.10 AM as SUMMA

We show now that the SUMMA class of sequential unconstrained minimization methods includes all the AM methods for which the five-point property holds.

### 9.10.1 Reformulating AM as SUMMA

For each $p$ in the set $P$, define $q(p)$ in $Q$ as a member of $Q$ for which $\Theta(p, q(p)) \leq \Theta(p, q)$, for all $q \in Q$. Let $f(p) = \Theta(p, q(p))$.

At the $n$th step of AM we minimize

$$G_n(p) = \Theta(p, q^{n-1}) = \Theta(p, q(p)) + \Big(\Theta(p, q^{n-1}) - \Theta(p, q(p))\Big) \tag{9.19}$$

to get $p^n$. With

$$g_n(p) = \Big(\Theta(p, q^{n-1}) - \Theta(p, q(p))\Big) \geq 0, \tag{9.20}$$

we can write

$$G_n(p) = f(p) + g_n(p). \tag{9.21}$$

According to the five-point property, we have

$$G_n(p) - G_n(p^n) \geq \Theta(p, q^n) - \Theta(p, q(p)) = g_{n+1}(p). \tag{9.22}$$

It follows that AM is a member of the SUMMA class.

## 9.11 SMART and EMML as SUMMA

We have seen that both the SMART and the EMML can be obtained as AM algorithms for which the five-point property holds. Consequently, both SMART and EMML are particular cases of SUMMA.

### 9.11.1 The SMART as SUMMA

In the case of SMART

$$\Theta(p, q) = KL(t(x), r(z)), \tag{9.23}$$

and

$$f(p) = \Theta(p, q(p)) = KL(t(x), r(x)) = KL(Px, y), \tag{9.24}$$

which is the function of $x$ we seek to minimize over $x \in \mathcal{X}$.

### 9.11.2 The EMML as SUMMA

In the case of EMML

$$\Theta(p,q) = KL(r(x), t(z)), \tag{9.25}$$

and

$$f(p) = \Theta(p, q(p)) = KL(r(x), t(x')), \tag{9.26}$$

which is not $KL(y, Px)$. In order to obtain the EMML from an AM formulation having the five-point property, and therefore to show that EMML is in the SUMMA class, we need to view the problem as minimizing not $KL(y, Px)$ but $f(x) = KL(r(x), t(x'))$. The minima are the same, however, as are the minimizers.

For the EMML we get $x^n = (x^{n-1})'$ by minimizing

$$G_n(x) = KL(r(x), t((x^{n-1})')) = f(x) + g_n(x), \tag{9.27}$$

where

$$g_n(x) = KL(r(x), t((x^{n-1})')) - KL(r(x), t(x')). \tag{9.28}$$

We need to show that

$$G_n(x) - G_n(x^n) \geq g_{n+1}(x). \tag{9.29}$$

From the Pythagorean identities for EMML we have

$$G_n(x) - G_n(x^n) = KL(r(x), r(x^n)), \tag{9.30}$$

and

$$g_{n+1}(x) = KL(x', (x^n)') \leq KL(r(x), r(x^n)), \tag{9.31}$$

which shows the EMML to be a member of the SUMMA class.

Consequently, the sequence $\{KL(y, Px^n)\}$ converges to the infimum of the function $KL(y, Px)$ over all $x \in \mathcal{X}$. The infimum is always attained at some $x \geq 0$ in the closure of $\mathcal{X}$ and it can be shown that the sequence $\{x^n\}$ converges to a minimizer of $KL(y, Px)$ over $x$ in the closure of $\mathcal{X}$ ([30, 32]).

## 9.12 Conclusion

It was shown previously in [46] that the SUMMA class includes a wide variety of optimization algorithms, including the barrier-function methods, the proximal minimization algorithm of Censor and Zenios [60, 61], the

entropic proximal method of Teboulle [148], and the simultaneous multiplicative algebraic reconstruction technique (SMART)[71, 138, 69, 30, 31]. With some reformulation, it also contains the penalty-function methods. We have now shown that the alternating minimization methods of [70] are included in the SUMMA class whenever the five-point property holds. As a consequence, we learn that the EMML algorithm for Poisson mixtures [139, 113, 149, 114, 30, 31] is also a member of the SUMMA class.

# Chapter 10

# The EM Algorithm

## 10.1 The Context

As I began studying the EMML algorithm for emission tomography, I was led naturally to the more general EM algorithm. The EM algorithm is not really a single algorithm, but a framework for the design of iterative likelihood maximization methods for parameter estimation; nevertheless, we shall continue to refer to *the* EM algorithm. The EM algorithm allows for both discrete-variable probability functions and continuous-variable probability density functions. The usual formulation of the EM algorithm is fine for the discrete case, but makes no sense in the continuous case. Nevertheless, most articles and books on the subject use this nonsensical formulation. I have tried for years to see how to replace it. Finally, in 2011, while working with Paul Eggermont, I hit on what I believe is the proper way to formulation the continuous case. This chapter describes that formulation.

## 10.2 Introduction

We suppose that the random vector $Y$ taking values in $R^N$ is governed by a probability density function (pdf) or probability function (pf) of the form $f_Y(y|\theta)$, for some value of the parameter vector $\theta \in \Theta$, where $\Theta$ is the set of all legitimate values of $\theta$. Our data consists of one realization $y$ of $Y$. The true vector of parameters is to be estimated by maximizing the likelihood function $L_y(\theta) = f_Y(y|\theta)$ over all $\theta \in \Theta$.

### 10.2.1 Simplifying the Computation

The basic idea underlying the EM algorithm is that there is another related random vector $X$, which we shall call the *preferred* data, such that, had

we been able to obtain one realization $x$ of $X$, maximizing the likelihood function $L_x(\theta) = f_X(x|\theta)$ would have been simpler than maximizing the likelihood function $L_y(\theta) = f_Y(y|\theta)$. In the missing-data model the preferred data is the *complete* data $X = Z = (Y, W)$, where $W$ is called the *missing* data.

The EM algorithm is not really a single algorithm, but a framework for the design of iterative likelihood maximization methods for parameter estimation; nevertheless, we shall continue to refer to *the* EM algorithm.

### 10.2.2 Missing Data

In the simplest version of the missing-data model, we assume that $M > N$ and that $Z = (Y, W)$, where $W$ is the missing-data random vector, taking values in $R^{M-N}$. If there were no missing data, we would have $z = (y, w)$, a realization of $Z$, and maximizing $L_z(\theta) = f(z|\theta)$ would be simpler. The choice of $W$ need not be unique.

More generally, the conventional formulation of the problem is that there is a random vector $X$ taking values in $R^M$, where $M \geq N$, with pdf or pf of the form $f_X(x|\theta)$, and a function $h : R^M \to R^N$ such that $Y = h(X)$. For example, let $X_1$ and $X_2$ be independent and uniformly distributed on $[0, \theta_0]$, $X = (X_1, X_2)$ and $Y = X_1 + X_2 = h(X)$. We can use the missing-data model here with $Z = (X_1 + X_2, X_1 - X_2)$. What is missing is not unique, however; instead of $W = X_1 - X_2$ as the missing data, we can use $W = X_2$, or any number of other combinations of $X_1$ and $X_2$ that would allow us to recapture $X$.

It is standard in the EM literature to call $X$ the complete data, $Y$ the incomplete data, and to assume that $Y = h(X)$ for some function $h$. This is because many, but not all, of the problems to which the EM approach is applied fit this description. As we shall attempt to convince the reader, this formulation is somewhat restrictive; the main point is simply that $f_X(x|\theta)$ would have been easier to maximize than $f_Y(y|\theta)$ is, regardless of the relationship between $X$ and $Y$. For this reason we shall call $Y$ the *given* data and $X$ the *preferred* data, and not assume that $Y = h(X)$ for some function $h$. We reserve the term *complete* data for $Z$ of the form $Z = (Y, W)$; note that, in this case we do have $Y = h(Z)$.

We shall assume, in all our theoretical discussions, that there is a likelihood maximizer $\theta_{ML}$ that maximizes the likelihood function $L_y(\theta)$ over $\theta \in \Theta$. In specific applications, the existence of a likelihood maximizer will depend on the problem.

In some applications, the preferred data $X$ arises naturally from the problem, while in other cases the user must imagine preferred data. This choice in selecting the preferred data can be helpful in speeding up the algorithm (see [83]).

### 10.2.3 A Multinomial Example

In many applications, the entries of the vector $y$ are independent realizations of a single real-valued or vector-valued random variable $V$, as they are, at least initially, for finite mixture problems to be considered later. This is not always the case, however, as the following example shows.

A well known example that was used in [74] and again in [120] to illustrate the EM algorithm concerns a multinomial model taken from genetics. Here there are four cells, with cell probabilities $\frac{1}{2}+\frac{1}{4}\theta_0$, $\frac{1}{4}(1-\theta_0)$, $\frac{1}{4}(1-\theta_0)$, and $\frac{1}{4}\theta_0$, for some $\theta_0 \in \Theta = [0,1]$ to be estimated. The entries of $y$ are the frequencies from a sample size of 197. We then have

$$f_Y(y|\theta) = \frac{197!}{y_1!y_2!y_3!y_4!}(\frac{1}{2}+\frac{1}{4}\theta)^{y_1}(\frac{1}{4}(1-\theta))^{y_2}(\frac{1}{4}(1-\theta))^{y_3}(\frac{1}{4}\theta)^{y_4}. \quad (10.1)$$

It is then supposed that the first of the original four cells can be split into two sub-cells, with probabilities $\frac{1}{2}$ and $\frac{1}{4}\theta_0$. We then write $y_1 = y_{11} + y_{12}$, and let

$$X = (Y_{11}, Y_{12}, Y_2, Y_3, Y_4),$$

where $X$ has a multinomial distribution with five cells. Note that we do now have $Y = h(X)$.

### 10.2.4 Difficulties with the Usual Formulation

In the literature on the EM algorithm, it is common to assume that there is a function $h : R^M \to R^N$ such that $Y = h(X)$. In the discrete case, in which summation and finite or infinite probability functions are involved, we then have

$$f_Y(y|\theta) = \sum_{x \in \mathcal{X}(y)} f_X(x|\theta), \quad (10.2)$$

where

$$\mathcal{X}(y) = \{x|h(x) = y\} = h^{-1}(\{y\}).$$

The difficulty arises in the continuous case, where integration and probability density functions (pdf) are needed; the set $\mathcal{X}(y)$ can have measure zero in $R^M$, so it is incorrect to mimic Equation (10.2) and write

$$f_Y(y|\theta) = \int_{x \in \mathcal{X}(y)} f_X(x|\theta)dx. \quad (10.3)$$

The case of $X_1$ and $X_2$ independent and uniformly distributed on $[0, \theta_0]$, $X = (X_1, X_2)$ and $Y = X_1 + X_2 = h(X)$ provides a good illustration of the problem. Here $\mathcal{X}(y)$ is the set of all pairs $(x_1, x_2)$ with $y = x_1 + x_2$; this subset of $R^2$ has Lebesgue measure zero in $R^2$. Now $f_{Y|X}(y|x, \theta)$ is a

delta function, which we may view as $\delta(x_2 - (y - x_1))$, with the property that

$$\int g(x_2)\delta(x_2 - (y - x_1))dx_2 = g(y - x_1).$$

Then we can write

$$f_Y(y|\theta) = \int\int \delta(x_2 - y - x_1) f_X(x_1, x_2|\theta)dx_2dx_1 = \int f_X(x_1, y - x_1)dx_1.$$

### 10.2.5 A Different Formulation

For any preferred data $X$ the EM algorithm involves two steps: given $y$ and the current estimate $\theta^k$, the E-step of the EM algorithm is to calculate

$$E(\log f_X(X|\theta)|y, \theta^k) = \int f_{X|Y}(x|y, \theta^k) \log f_X(x|\theta)dx, \tag{10.4}$$

the conditional expected value of the random vector $\log f_X(X|\theta)$, given $y$ and $\theta^k$. Then the M-step is to maximize

$$\int f_{X|Y}(x|y, \theta^k) \log f_X(x|\theta)dx \tag{10.5}$$

with respect to $\theta \in \Theta$ to obtain $\theta^{k+1}$. For the missing-data model $Z = (Y, W)$, we shall show that the M-step is to maximize

$$\int f_{W|Y}(w|y, \theta^k) \log f_{Y,W}(y, w|\theta)dw. \tag{10.6}$$

We shall also show that for the missing-data model we always have $L_y(\theta^{k+1}) \geq L_y(\theta^k)$.

This suggests that, for arbitrary preferred data $X$, we view $X$ as $W$, the missing data, and use $Z = (Y, X)$. When this works, the likelihood will be non-decreasing. As we shall see, this approach does work if the preferred data $X$ satisfies the *acceptability condition* $f_{Y|X}(y|x, \theta) = f_{Y|X}(y|x)$; that is, the conditional distribution of $Y$, given $X$, exists and is independent of the parameter vector $\theta$. Consequently, whenever the preferred data $X$ is acceptable, we know that $L_y(\theta^{k+1}) \geq L_y(\theta^k)$.

For those cases involving continuous distributions in which $Y = h(X)$ the $X$ is acceptable, but we must define $f_{Y|X}(y|x)$ in terms of a delta function; we shall not consider such cases here.

### 10.2.6 The Example of Finite Mixtures

We say that a random vector $V$ taking values in $R^M$ is a *finite mixture* if, for $j = 1, ..., J$, $f_j$ is a probability density function or probability function,

$\theta_j \geq 0$ is a weight, the $\theta_j$ sum to one, and the probability density function or probability function for $V$ is

$$f_V(v|\theta) = \sum_{j=1}^{J} \theta_j f_j(v). \tag{10.7}$$

We draw $N$ independent samples of $V$, denoted $v^n$, and let $y^n$, the $n$th entry of the vector $y$ be the vector $v^n$. To create the preferred data we assume that, for each $n$, the vector $v^n$ is a sample of the random vector $V^n$ whose pdf or pf is $f_{j_n}$, where the probability that $j_n = j$ is $\theta_j$. We then let the $N$ entries of the preferred data $X$ be the indices $j_n$. The conditional distribution of $Y$, given $X$, clearly is independent of the parameter vector $\theta$, and is given by

$$f_Y(y|x,\theta) = \prod_{n=1}^{N} f_{j_n}(y_n).$$

Therefore, $X$ is acceptable. Note that we cannot recapture the entries of $y$ from those of $x$, so the model $Y = h(X)$ does not hold here. Note also that, although the vector $y$ is taken originally to be a vector whose entries are independently drawn samples from $V$, when we create the preferred data $X$ we change our view of $y$. Now each entry of $y$ is governed by a different distribution, so $y$ is no longer viewed as a vector of independent sample values of a single random vector.

### 10.2.7 Overview

We begin by considering in detail the missing-data model. The EM algorithm we obtain there leads to non-decreasing likelihood. When we attempt to treat arbitrary preferred data $X$ as if it were missing data, to take advantage of the non-decreasing likelihood, we find that this approach works, provided $X$ satisfies an acceptability condition. Acceptability also permits the reformulation of the EM algorithm in terms of alternating minimization of a Kullback-Leibler distance, along the lines of the work of Csiszár and Tusnády.

We turn then to several examples. The first is the sum of independent Poisson random variables and its application in emission tomography, leading to a special case of the EM algorithm that we call here the EMML algorithm. Next, we derive the Mix-EM algorithm for finite mixture problems. Because the preferred data is acceptable, the likelihood is non-decreasing. Going further, we use a convergence theorem obtained elsewhere for emission tomography to prove convergence of the Mix-EM algorithm to a vector of mixing proportions that is a likelihood maximizer.

## 10.3 The Missing-Data Model

For any measurable function $H(y, w)$ we have

$$E(H(Y,W)|\theta^k) = \int\int H(y,w) f_{Y,W}(y,w|\theta^k) dw\, dy, \tag{10.8}$$

so that

$$E(H(Y,W)|\theta^k) = \int \Big( \int H(y,w) \frac{f_{Y,W}(y,w|\theta^k)}{f_Y(y|\theta^k)} dw \Big) f_Y(y|\theta^k) dy. \tag{10.9}$$

We also have

$$E(H(Y,W)|\theta^k) = \int \Big( E(H(Y,W)|y,\theta^k) \Big) f_Y(y|\theta^k) dy, \tag{10.10}$$

from which we conclude that

$$\int H(y,w) \frac{f_{Y,W}(y,w|\theta^k)}{f_Y(y|\theta^k)} dw = E(H(Y,W)|y,\theta^k) = \int H(y,w) f_{W|Y}(w|y,\theta^k) dw. \tag{10.11}$$

Substituting $\log f_Z(Z|\theta) = \log f_{Y,W}(Y,W|\theta)$ for $H(Y,W)$, the E-step of the EM algorithm becomes

$$E(\log f_Z(Z|\theta)|y,\theta^k) = \int \log f_{Y,W}(y,w|\theta) f_{W|Y}(w|y,\theta^k) dw. \tag{10.12}$$

Then the M-step is to maximize

$$\int \log f_{Y,W}(y,w|\theta) f_{W|Y}(w|y,\theta^k) dw \tag{10.13}$$

to get $\theta^{k+1}$.

For the missing-data model we have the following result (see [101, 53]).

**Proposition 10.1** *The sequence $\{L_y(\theta^k)\}$ is non-decreasing, as $k \to +\infty$.*

**Proof:** Begin with

$$\log f_Y(y|\theta) = \int f_{W|Y}(w|y,\theta^k) \log f_Y(y|\theta) dw.$$

Then

$$\log f_Y(y|\theta) = \int f_{W|Y}(w|y,\theta^k) \Big( \log f_{Y,W}(y,w|\theta) - \log f_{W|Y}(w|y,\theta) \Big) dw.$$

Therefore,

$$\log f_Y(y|\theta^{k+1}) - \log f_Y(y|\theta^k) =$$

$$\int f_{W|Y}(w|y,\theta^k)\log f_{Y,W}(y,w|\theta^{k+1})dw - \int f_{W|Y}(w|y,\theta^k)\log f_{Y,W}(y,w|\theta^k)dw +$$

$$\int f_{W|Y}(w|y,\theta^k)\log f_{W|Y}(w|y,\theta^k)dw - \int f_{W|Y}(w|y,\theta^k)\log f_{W|Y}(w|y,\theta^{k+1})dw. \quad (10.14)$$

The first difference on the right side of Equation (10.14) is non-negative because of the M-step, while the second difference is non-negative because of Jensen's Inequality. ∎

The fact that likelihood is not decreasing in the missing-data model suggests that we try to use this model for arbitrary preferred data $X$, by defining the complete data to be $Z = (Y, X)$ and viewing $X$ as the missing data, that is, $W = X$. This approach works, provided that $X$ satisfies the acceptability condition.

## 10.4 The EM Algorithm for Acceptable $X$

For any preferred data $X$ the E-step of the EM algorithm is to calculate

$$E(\log f_X(X|\theta)|y,\theta^k) = \int f_{X|Y}(x|y,\theta^k)\log f_X(x|\theta)dx. \quad (10.15)$$

Once we have $y$, the M-step is then to maximize

$$E(\log f_X(X|\theta)|y,\theta^k) = \int f_{X|Y}(x|y,\theta^k)\log f_X(x|\theta)dx \quad (10.16)$$

to get $\theta = \theta^{k+1}$.

### 10.4.1 The Likelihood is Non-Decreasing

For the moment, let $X$ be any preferred data. We examine the behavior of the likelihood when the EM algorithm is applied to $X$.

As in the proof for the missing-data model, we begin with

$$\log f_Y(y|\theta) = \int f_{X|Y}(x|y,\theta^k)\log f_Y(y|\theta)dx.$$

Then

$$\log f_Y(y|\theta) = \int f_{X|Y}(x|y,\theta^k)\Big(\log f_{Y,X}(y,x|\theta) - \log f_{X|Y}(x|y,\theta)\Big)dx.$$

Therefore,

$$\log f_Y(y|\theta^{k+1}) - \log f_Y(y|\theta^k) =$$

$$\int f_{X|Y}(x|y,\theta^k)\log f_{Y,X}(y,x|\theta^{k+1})dx - \int f_{X|Y}(x|y,\theta^k)\log f_{Y,X}(y,x|\theta^k)dx +$$

$$\int f_{X|Y}(x|y,\theta^k)\log f_{X|Y}(x|y,\theta^k)dx - \int f_{X|Y}(x|y,\theta^k)\log f_{X|Y}(x|y,\theta^{k+1})dx \quad (10.17)$$

The second difference on the right side of Equation (10.17) is non-negative because of Jensen's Inequality. But we cannot assert that the first difference on the right is non-negative because, in the M-step, we maximize

$$\int f_{X|Y}(x|y,\theta^k)\log f_X(x|\theta)dx$$

not

$$\int f_{X|Y}(x|y,\theta^k)\log f_{Y,X}(y,x|\theta)dx.$$

If $X$ is acceptable, then

$$\log f_{Y,X}(y,x|\theta) - \log f_X(x|\theta) = \log f_{Y|X}(y|x)$$

is independent of $\theta$ and the difficulty disappears. We may then conclude that likelihood is non-decreasing for acceptable $X$.

### 10.4.2 Generalized EM Algorithms

If, instead of maximizing

$$\int f_{X|Y}(x|y,\theta^k)\log f_X(x|\theta)dx,$$

at each M-step, we simply select $\theta^{k+1}$ so that

$$\int f_{X|Y}(x|y,\theta^k)\log f_{Y,X}(y,x|\theta^{k+1})dx - \int f_{X|Y}(x|y,\theta^k)\log f_{Y,X}(y,x|\theta^k)dx > 0,$$

we say that we are using a *generalized* EM (GEM) algorithm. It is clear from the discussion in the previous subsection that, whenever $X$ is acceptable, a GEM also guarantees that likelihood is non-decreasing.

### 10.4.3 Preferred Data as Missing Data

We know that, when the missing-data model is used and the M-step is defined as maximizing the function in (10.13), the likelihood is not decreasing. It would seem then that, for any choice of preferred data $X$, we could view this data as missing and take as our complete data the pair $Z = (Y, X)$, with $X$ now playing the role of $W$. The function in (10.13) is then

$$\int f_{X|Y}(x|y,\theta^k)\log f_{Y,X}(y,x|\theta)dx; \quad (10.18)$$

we maximize this function to get $\theta^{k+1}$. It then follows that $L_y(\theta^{k+1}) \geq L_y(\theta^k)$. The obvious question is whether or not these two functions given in (10.15) and (10.18) have the same maximizers.

For acceptable $X$ we have

$$\log f_{Y,X}(y, x|\theta) = \log f_X(x|\theta) + \log f_{Y|X}(y|x), \tag{10.19}$$

so the two functions given in (10.15) and (10.18) do have the same maximizers. It follows once again that, whenever the preferred data is acceptable, we have $L_y(\theta^{k+1}) \geq L_y(\theta^k)$. Without additional assumptions, however, we cannot conclude that $\{\theta^k\}$ converges to $\theta_{ML}$, nor that $\{f_Y(y|\theta^k)\}$ converges to $f_Y(y|\theta_{ML})$.

In the discrete case in which $Y = h(X)$ the conditional probability $f_{Y|X}(y|x, \theta)$ is $\delta(y - h(x))$, as a function of $y$, for given $x$, and is the characteristic function of the set $\mathcal{X}(y)$, as a function of $x$, for given $y$. Therefore, we can write

$$f_{X|Y}(x|y, \theta) = \begin{cases} f_X(x|\theta)/f_Y(y|\theta), & \text{if } x \in \mathcal{X}(y); \\ 0, & \text{if } x \notin \mathcal{X}(y). \end{cases} \tag{10.20}$$

For the continuous case in which $Y = h(X)$, the pdf $f_{Y|X}(y|x, \theta)$ is again a delta function of $y$, for given $x$; the difficulty arises when we need to view this as a function of $x$, for given $y$. The acceptability property helps us avoid this difficulty.

When $X$ is acceptable, we have

$$f_{X|Y}(x|y, \theta) = f_{Y|X}(y|x) f_X(x|\theta)/f_Y(y|\theta), \tag{10.21}$$

whenever $f_Y(y|\theta) \neq 0$, and is zero otherwise. Consequently, when $X$ is acceptable, we have a kernel model for $f_Y(y|\theta)$ in terms of the $f_X(x|\theta)$:

$$f_Y(y|\theta) = \int f_{Y|X}(y|x) f_X(x|\theta) dx; \tag{10.22}$$

for the continuous case we view this as a corrected version of Equation (10.3). In the discrete case the integral is replaced by a summation, of course, but when we are speaking generally about either case, we shall use the integral sign.

The acceptability of the missing data $W$ is used in [53], but more for computational convenience and to involve the Kullback-Leibler distance in the formulation of the EM algorithm. It is not necessary that $W$ be acceptable in order for likelihood to be non-decreasing, as we have seen.

## 10.5 The EM and the Kullback-Leibler Distance

We illustrate the usefulness of acceptability and reformulate the M-step in terms of cross-entropy or Kullback-Leibler distance minimization.

### 10.5.1 Cross-Entropy or the Kullback-Leibler Distance

The cross-entropy or Kullback-Leibler distance is a useful tool for analyzing the EM algorithm. For positive numbers $u$ and $v$, the Kullback-Leibler distance from $u$ to $v$ is

$$KL(u,v) = u\log\frac{u}{v} + v - u. \tag{10.23}$$

We also define $KL(0,0) = 0$, $KL(0,v) = v$ and $KL(u,0) = +\infty$. The KL distance is extended to nonnegative vectors component-wise, so that for nonnegative vectors $a$ and $b$ we have

$$KL(a,b) = \sum_{j=1}^{J} KL(a_j, b_j). \tag{10.24}$$

One of the most useful facts about the KL distance is contained in the following lemma.

**Lemma 10.1** *For non-negative vectors $a$ and $b$, with $b_+ = \sum_{j=1}^{J} b_j > 0$, we have*

$$KL(a,b) = KL(a_+, b_+) + KL(a, \frac{a_+}{b_+}b). \tag{10.25}$$

### 10.5.2 Using Acceptable Data

The assumption that the data $X$ is acceptable helps simplify the theoretical discussion of the EM algorithm.

For any preferred $X$ the M-step of the EM algorithm, in the continuous case, is to maximize the function

$$\int f_{X|Y}(x|y,\theta^k)\log f_X(x|\theta)dx, \tag{10.26}$$

over $\theta \in \Theta$; the integral is replaced by a sum in the discrete case. For notational convenience we let

$$b(\theta^k) = f_{X|Y}(x|y,\theta^k), \tag{10.27}$$

and

$$f(\theta) = f_X(x|\theta). \tag{10.28}$$

Then the M-step is equivalent to minimizing the Kullback-Leibler or cross-entropy distance

$$KL(b(\theta^k), f(\theta)) = \int f_{X|Y}(x|y,\theta^k)\log\Big(\frac{f_{X|Y}(x|y,\theta^k)}{f_X(x|\theta)}\Big)dx$$

$$= \int f_{X|Y}(x|y,\theta^k) \log\Big(\frac{f_{X|Y}(x|y,\theta^k)}{f_X(x|\theta)}\Big) + f_X(x|\theta) - f_{X|Y}(x|y,\theta^k)dx \quad (10.29)$$

This holds since both $f_X(x|\theta)$ and $f_{X|Y}(x|y,\theta^k)$ are probability density functions or probabilities.

For acceptable $X$ we have

$$\log f_{Y,X}(y,x|\theta) = \log f_{X|Y}(x|y,\theta) + \log f_Y(y|\theta) = \log f_{Y|X}(y|x) + \log f_X(x|\theta) \quad (10.30)$$

Therefore,

$$\log f_Y(y|\theta^{k+1}) - \log f_Y(y|\theta) =$$

$$KL(b(\theta^k), f(\theta)) - KL(b(\theta^k), f(\theta^{k+1})) + KL(b(\theta^k), b(\theta^{k+1})) - KL(b(\theta^k), b(\theta)) \quad (10.31)$$

Since $\theta = \theta^{k+1}$ minimizes $KL(b(\theta^k), f(\theta))$, we have that

$$\log f_Y(y|\theta^{k+1}) - \log f_Y(y|\theta^k) =$$

$$KL(b(\theta^k), f(\theta^k)) - KL(b(\theta^k), f(\theta^{k+1})) + KL(b(\theta^k), b(\theta^{k+1})) \geq 0. \quad (10.32)$$

This tells us, once again, that the sequence of likelihood values $\{\log f_Y(y|\theta^k)\}$ is increasing, and that the sequence of its negatives, $\{-\log f_Y(y|\theta^k)\}$, is decreasing. Since we assume that there is a maximizer $\theta_{ML}$ of the likelihood, the sequence $\{-\log f_Y(y|\theta^k)\}$ is also bounded below and the sequences $\{KL(b(\theta^k), b(\theta^{k+1}))\}$ and $\{KL(b(\theta^k), f(\theta^k)) - KL(b(\theta^k), f(\theta^{k+1}))\}$ converge to zero.

Without some notion of convergence in the parameter space $\Theta$, we cannot conclude that $\{\theta^k\}$ converges to a maximum likelihood estimate $\theta_{ML}$. Without some additional assumptions, we cannot even conclude that the functions $f(\theta^k)$ converge to $f(\theta_{ML})$.

## 10.6 The Approach of Csiszár and Tusnády

For acceptable $X$ the M-step of the EM algorithm is to minimize the function $KL(b(\theta^k), f(\theta))$ over $\theta \in \Theta$ to get $\theta^{k+1}$. To put the EM algorithm into the framework of the *alternating minimization* approach of Csiszár and Tusnády [70], we need to view the M-step in a slightly different way; the problem is that, for the continuous case, having found $\theta^{k+1}$, we do not then minimize $KL(b(\theta), f(\theta^{k+1}))$ at the next step.

### 10.6.1 The Framework of Csiszár and Tusnády

Following [70], we take $\Psi(p,q)$ to be a real-valued function of the variables $p \in P$ and $q \in Q$, where $P$ and $Q$ are arbitrary sets. Minimizing $\Psi(p, q^n)$ gives $p^{n+1}$ and minimizing $\Psi(p^{n+1}, q)$ gives $q^{n+1}$, so that

$$\Psi(p^n, q^n) \geq \Psi(p^n, q^{n+1}) \geq \Psi(p^{n+1}, q^{n+1}). \quad (10.33)$$

The objective is to find $(\hat{p}, \hat{q})$ such that

$$\Psi(p, q) \geq \Psi(\hat{p}, \hat{q}),$$

for all $p$ and $q$. In order to show that $\{\Psi(p^n, q^n)\}$ converges to

$$d = \inf_{p \in P, q \in Q} \Psi(p, q)$$

the authors of [70] assume the three- and four-point properties.

If there is a non-negative function $\Delta : P \times P \rightarrow R$ such that

$$\Psi(p, q^{n+1}) - \Psi(p^{n+1}, q^{n+1}) \geq \Delta(p, p^{n+1}), \tag{10.34}$$

then the *three-point property* holds. If

$$\Delta(p, p^n) + \Psi(p, q) \geq \Psi(p, q^{n+1}), \tag{10.35}$$

for all $p$ and $q$, then the *four-point property* holds. Combining these two inequalities, we have

$$\Delta(p, p^n) - \Delta(p, p^{n+1}) \geq \Psi(p^{n+1}, q^{n+1}) - \Psi(p, q). \tag{10.36}$$

From the inequality in (10.36) it follows easily that the sequence $\{\Psi(p^n, q^n)\}$ converges to $d$. Suppose this is not the case. Then there are $p'$, $q'$, and $D > d$ with

$$\Psi(p^n, q^n) \geq D > \Psi(p', q') \geq d.$$

From Equation (10.36) we have

$$\Delta(p', p^n) - \Delta(p', p^{n+1}) \geq \Psi(p^{n+1}, q^{n+1}) - \Psi(p', q') \geq D - d > 0.$$

But since $\{\Delta(p', p^n)\}$ is a decreasing sequence of positive quantities, successive differences must converge to zero.

### 10.6.2 Alternating Minimization for the EM

Assume that $X$ is acceptable. We define the function $G(\theta)$ to be

$$G(\theta) = \int f_{X|Y}(x|y, \theta) \log f_{Y|X}(y|x) dx, \tag{10.37}$$

for the continuous case, with a sum replacing the integral for the discrete case. Using the identities

$$f_{Y,X}(y, x|\theta) = f_{X|Y}(x|y, \theta) f_Y(y|\theta) = f_{Y|X}(y|x, \theta) f_X(x|\theta) = f_{Y|X}(y|x) f_X(x|\theta),$$

we then have

$$\log f_Y(y|\theta) = G(\theta') + KL(b(\theta'), b(\theta)) - KL(b(\theta'), f(\theta)), \tag{10.38}$$

for any parameter values $\theta$ and $\theta'$. With the choice of $\theta' = \theta$ we have

$$\log f_Y(y|\theta) = G(\theta) - KL(b(\theta), f(\theta)). \tag{10.39}$$

Therefore, subtracting Equation 10.39 from Equation 10.38, we get

$$\Big(KL(b(\theta'), f(\theta)) - G(\theta')\Big) - \Big(KL(b(\theta), f(\theta)) - G(\theta)\Big) = KL(b(\theta'), b(\theta)). \tag{10.40}$$

Now we can put the EM algorithm into the alternating-minimization framework.

Define

$$\Psi(b(\theta'), f(\theta)) = KL(b(\theta'), f(\theta)) - G(\theta'). \tag{10.41}$$

We know from Equation (10.40) that

$$\Psi(b(\theta'), f(\theta)) - \Psi(b(\theta), f(\theta)) = KL(b(\theta'), b(\theta)). \tag{10.42}$$

Therefore, we can say that the M-step of the EM algorithm is to minimize $\Psi(b(\theta^k), f(\theta))$ over $\theta \in \Theta$ to get $\theta^{k+1}$ and that minimizing $\Psi(b(\theta), f(\theta^{k+1}))$ gives us $\theta = \theta^{k+1}$ again. With the choice of

$$\Delta(b(\theta'), b(\theta)) = KL(b(\theta'), b(\theta)),$$

Equation (10.42) becomes

$$\Psi(b(\theta'), f(\theta)) - \Psi(b(\theta), f(\theta)) = \Delta(b(\theta'), b(\theta)), \tag{10.43}$$

which is the three-point property.

With $P = \mathcal{B}(\Theta)$ and $Q = \mathcal{F}(\Theta)$ the collections of all functions $b(\theta)$ and $f(\theta)$, respectively, we can view the EM algorithm as alternating minimization of the function $\Psi(p, q)$, over $p \in P$ and $q \in Q$. As we have seen, the three-point property holds. What about the four-point property?

The Kullback-Leibler distance is an example of a jointly convex Bregman distance. According to a lemma of Eggermont and LaRiccia [78, 79], the four-point property holds for alternating minimization of such distances, provided that the objects that can occur in the second-variable position form a convex subset of $R^N$. In the continuous case of the EM algorithm, we are not performing alternating minimization on the function $KL(b(\theta), f(\theta'))$, but on $KL(b(\theta), f(\theta')) + G(\theta)$. In the discrete case, whenever $Y = h(X)$, the function $G(\theta)$ is always zero, so we are performing alternating minimization on the KL distance $KL(b(\theta), f(\theta'))$. In [11] the authors consider the problem of minimizing a function of the form

$$\Lambda(p, q) = \phi(p) + \psi(q) + D_f(p, q), \tag{10.44}$$

where $\phi$ and $\psi$ are convex and differentiable on $R^J$, $D_f$ is a Bregman distance, and $P = Q$ is the interior of the domain of $f$. In [52] it was shown that, when $D_f$ is jointly convex, the function $\Lambda(p, q)$ has the five-point property of [70]. In some particular instances, the collection of the functions $f(\theta)$ is a convex set, as well, so the three- and four-point properties hold.

## 10.7 Sums of Independent Poisson Random Variables

The EM is often used with aggregated data. The case of sums of independent Poisson random variables is particularly important.

### 10.7.1 Poisson Sums

Let $X_1, ..., X_N$ be independent Poisson random variables with expected value $E(X_n) = \lambda_n$. Let $X$ be the random vector with $X_n$ as its entries, $\lambda$ the vector whose entries are the $\lambda_n$, and $\lambda_+ = \sum_{n=1}^N \lambda_n$. Then the probability function for $X$ is

$$f_X(x|\lambda) = \prod_{n=1}^N \lambda_n^{x_n} \exp(-\lambda_n)/x_n! = \exp(-\lambda_+) \prod_{n=1}^N \lambda_n^{x_n}/x_n! \,. \quad (10.45)$$

Now let $Y = \sum_{n=1}^N X_n$. Then, the probability function for $Y$ is

$$\begin{aligned} \text{Prob}(Y = y) &= \text{Prob}(X_1 + ... + X_N = y) \\ &= \sum_{x_1+...x_N=y} \exp(-\lambda_+) \prod_{n=1}^N \lambda_n^{x_n}/x_n! \,. \end{aligned} \quad (10.46)$$

As we shall see shortly, we have

$$\sum_{x_1+...x_N=y} \exp(-\lambda_+) \prod_{n=1}^N \lambda_n^{x_n}/x_n! = \exp(-\lambda_+)\lambda_+^y/y! \,. \quad (10.47)$$

Therefore, $Y$ is a Poisson random variable with $E(Y) = \lambda_+$.

When we observe an instance of $Y$, we can consider the conditional distribution $f_{X|Y}(x|y, \lambda)$ of $\{X_1, ..., X_N\}$, subject to $y = X_1 + ... + X_N$. We have

$$f_{X|Y}(x|y, \lambda) = \frac{y!}{x_1!...x_N!} \Big(\frac{\lambda_1}{\lambda_+}\Big)^{x_1} ... \Big(\frac{\lambda_N}{\lambda_+}\Big)^{x_N}. \quad (10.48)$$

This is a *multinomial distribution*.

Given $y$ and $\lambda$, the conditional expected value of $X_n$ is then

$$E(X_n|y, \lambda) = y\lambda_n/\lambda_+.$$

To see why this is true, consider the marginal conditional distribution $f_{X_1|Y}(x_1|y, \lambda)$ of $X_1$, conditioned on $y$ and $\lambda$, which we obtain by holding $x_1$ fixed and summing over the remaining variables. We have

$$f_{X_1|Y}(x_1|y, \lambda) = \frac{y!}{x_1!(y - x_1)!} \Big(\frac{\lambda_1}{\lambda_+}\Big)^{x_1} \Big(\frac{\lambda'_+}{\lambda_+}\Big)^{y-x_1} \sum_{x_2+...+x_N=y-x_1} \frac{(y - x_1)!}{x_2!...x_N!} \prod_{n=2}^N \Big(\frac{\lambda_n}{\lambda'_+}\Big)^{x_n},$$

where

$$\lambda'_+ = \lambda_+ - \lambda_1.$$

As we shall show shortly,

$$\sum_{x_2+...+x_N=y-x_1} \frac{(y-x_1)!}{x_2!...x_N!} \prod_{n=2}^{N} \Big(\frac{\lambda_n}{\lambda'_+}\Big)^{x_n} = 1,$$

so that

$$f_{X_1|Y}(x_1|y,\lambda) = \frac{y!}{x_1!(y-x_1)!}\Big(\frac{\lambda_1}{\lambda_+}\Big)^{x_1}\Big(\frac{\lambda'_+}{\lambda_+}\Big)^{y-x_1}.$$

The random variable $X_1$ is equivalent to the random number of heads showing in $y$ flips of a coin, with the probability of heads given by $\lambda_1/\lambda_+$. Consequently, the conditional expected value of $X_1$ is $y\lambda_1/\lambda_+$, as claimed. In the next subsection we look more closely at the multinomial distribution.

### 10.7.2 The Multinomial Distribution

When we expand the quantity $(a_1 + ... + a_N)^y$, we obtain a sum of terms, each having the form $a_1^{x_1}...a_N^{x_N}$, with $x_1 + ... + x_N = y$. How many terms of the same form are there? There are $N$ variables $a_n$. We are to use $x_n$ of the $a_n$, for each $n = 1, ..., N$, to get $y = x_1 + ... + x_N$ factors. Imagine $y$ blank spaces, each to be filled in by a variable as we do the selection. We select $x_1$ of these blanks and mark them $a_1$. We can do that in $\binom{y}{x_1}$ ways. We then select $x_2$ of the remaining blank spaces and enter $a_2$ in them; we can do this in $\binom{y-x_1}{x_2}$ ways. Continuing in this way, we find that we can select the $N$ factor types in

$$\binom{y}{x_1}\binom{y-x_1}{x_2}...\binom{y-(x_1+...+x_{N-2})}{x_{N-1}} \tag{10.49}$$

ways, or in

$$\frac{y!}{x_1!(y-x_1)!}...\frac{(y-(x_1+...+x_{N-2}))!}{x_{N-1}!(y-(x_1+...+x_{N-1}))!} = \frac{y!}{x_1!...x_N!}. \tag{10.50}$$

This tells us in how many different sequences the factor variables can be selected. Applying this, we get the multinomial theorem:

$$(a_1 + ... + a_N)^y = \sum_{x_1+...+x_N=y} \frac{y!}{x_1!...x_N!} a_1^{x_1}...a_N^{x_N}. \tag{10.51}$$

Select $a_n = \lambda_n/\lambda_+$. Then,

$$1 = 1^y = \Big(\frac{\lambda_1}{\lambda_+} + ... + \frac{\lambda_N}{\lambda_+}\Big)^y$$

$$= \sum_{x_1+\ldots+x_N=y} \frac{y!}{x_1!\ldots x_N!}\Big(\frac{\lambda_1}{\lambda_+}\Big)^{x_1}\ldots\Big(\frac{\lambda_N}{\lambda_+}\Big)^{x_N}. \tag{10.52}$$

From this we get

$$\sum_{x_1+\ldots x_N=y} \exp(-\lambda_+)\prod_{n=1}^{N}\lambda_n^{x_n}/x_n! = \exp(-\lambda_+)\lambda_+^y/y!\ . \tag{10.53}$$

## 10.8 Poisson Sums in Emission Tomography

Sums of Poisson random variables and the problem of complete versus incomplete data arise in *single-photon computed emission tomography* (SPECT) (Wernick and Aarsvold (2004) [150]).

### 10.8.1 The SPECT Reconstruction Problem

In their 1976 paper Rockmore and Makovski [137] suggested that the problem of reconstructing a tomographic image be viewed as statistical parameter estimation. Shepp and Vardi (1982) [139] expanded on this idea and suggested that the EM algorithm discussed by Dempster, Laird and Rubin (1977) [74] be used for the reconstruction. The region of interest within the body of the patient is discretized into $J$ pixels (or voxels), with $\lambda_j \geq 0$ the unknown amount of radionuclide within the $j$th pixel; we assume that $\lambda_j$ is also the expected number of photons emitted from the $j$th pixel during the scanning time. Emitted photons are detected at any one of $I$ detectors outside the body, with $y_i > 0$ the photon count at the $i$th detector. The probability that a photon emitted at the $j$th pixel will be detected at the $i$th detector is $P_{ij}$, which we assume is known; the overall probability of detecting a photon emitted from the $j$th pixel is $s_j = \sum_{i=1}^{I} P_{ij} > 0$.

#### The Preferred Data

For each $i$ and $j$ the random variable $X_{ij}$ is the number of photons emitted from the $j$th pixel and detected at the $i$th detector; the $X_{ij}$ are assumed to be independent and $P_{ij}\lambda_j$-Poisson. With $x_{ij}$ a realization of $X_{ij}$, the vector $x$ with components $x_{ij}$ is our preferred data. The pdf for this preferred data is a probability vector, with

$$f_X(x|\lambda) = \prod_{i=1}^{I}\prod_{j=1}^{J}\exp^{-P_{ij}\lambda_j}(P_{ij}\lambda_j)^{x_{ij}}/x_{ij}!\ . \tag{10.54}$$

Given an estimate $\lambda^k$ of the vector $\lambda$ and the restriction that $Y_i = \sum_{j=1}^{J} X_{ij}$, the random variables $X_{i1}, \ldots, X_{iJ}$ have the multinomial distri-

bution

$$\text{Prob}(x_{i1}, ..., x_{iJ}) = \frac{y_i!}{x_{i1}! \cdots x_{iJ}!} \prod_{j=1}^{J} \Big( \frac{P_{ij}\lambda_j}{(P\lambda)_i} \Big)^{x_{ij}}.$$

Therefore, the conditional expected value of $X_{ij}$, given $y$ and $\lambda^k$, is

$$E(X_{ij}|y, \lambda^k) = \lambda_j^k P_{ij} \Big( \frac{y_i}{(P\lambda^k)_i} \Big),$$

and the conditional expected value of the random variable

$$\log f_X(X|\lambda) = \sum_{i=1}^{I} \sum_{j=1}^{J} (-P_{ij}\lambda_j) + X_{ij} \log(P_{ij}\lambda_j) + \text{constants}$$

becomes

$$E(\log f_X(X|\lambda)|y, \lambda^k) = \sum_{i=1}^{I} \sum_{j=1}^{J} \left( (-P_{ij}\lambda_j) + \lambda_j^k P_{ij} \Big( \frac{y_i}{(P\lambda^k)_i} \Big) \log(P_{ij}\lambda_j) \right),$$

omitting terms that do not involve the parameter vector $\lambda$. In the EM algorithm, we obtain the next estimate $\lambda^{k+1}$ by maximizing $E(\log f_X(X|\lambda)|y, \lambda^k)$.

The log likelihood function for the preferred data $X$ (omitting constants) is

$$LL_x(\lambda) = \sum_{i=1}^{I} \sum_{j=1}^{J} \Big( - P_{ij}\lambda_j + X_{ij} \log(P_{ij}\lambda_j) \Big). \tag{10.55}$$

Of course, we do not have the complete data.

**The Incomplete Data**

What we do have are the $y_i$, values of the random variables

$$Y_i = \sum_{j=1}^{J} X_{ij}; \tag{10.56}$$

this is the given data. These random variables are also independent and $(P\lambda)_i$-Poisson, where

$$(P\lambda)_i = \sum_{j=1}^{J} P_{ij}\lambda_j.$$

The log likelihood function for the given data is

$$LL_y(\lambda) = \sum_{i=1}^{I} \Big( - (P\lambda)_i + y_i \log((P\lambda)_i) \Big). \tag{10.57}$$

Maximizing $LL_x(\lambda)$ in Equation (10.55) is easy, while maximizing $LL_y(\lambda)$ in Equation (10.57) is harder and requires an iterative method.

The EM algorithm involves two steps: in the E-step we compute the conditional expected value of $LL_x(\lambda)$, conditioned on the data vector $y$ and the current estimate $\lambda^k$ of $\lambda$; in the M-step we maximize this conditional expected value to get the next $\lambda^{k+1}$. Putting these two steps together, we have the following EMML iteration:

$$\lambda_j^{k+1} = \lambda_j^k s_j^{-1} \sum_{i=1}^{I} P_{ij} \frac{y_i}{(P\lambda^k)_i}. \tag{10.58}$$

For any positive starting vector $\lambda^0$, the sequence $\{\lambda^k\}$ converges to a maximizer of $LL_y(\lambda)$, over all non-negative vectors $\lambda$.

Note that, because we are dealing with finite probability vectors in this example, it is a simple matter to conclude that

$$f_Y(y|\lambda) = \sum_{x \in \mathcal{X}(y)} f_X(x|\lambda). \tag{10.59}$$

### 10.8.2 Using the KL Distance

In this subsection we assume, for notational convenience, that the system $y = P\lambda$ has been normalized so that $s_j = 1$ for each $j$. Maximizing $E(\log f_X(X|\lambda)|y, \lambda^k)$ is equivalent to minimizing $KL(r(\lambda^k), q(\lambda))$, where $r(\lambda)$ and $q(\lambda)$ are $I$ by $J$ arrays with entries

$$r(\lambda)_{ij} = \lambda_j P_{ij} \Big( \frac{y_i}{(P\lambda)_i} \Big),$$

and

$$q(\lambda)_{ij} = \lambda_j P_{ij}.$$

In terms of our previous notation we identify $r(\lambda)$ with $b(\theta)$, and $q(\lambda)$ with $f(\theta)$. The set $\mathcal{F}(\Theta)$ of all $f(\theta)$ is now a convex set and the four-point property of [70] holds. The iterative step of the EMML algorithm is then

$$\lambda_j^{k+1} = \lambda_j^k \sum_{i=1}^{I} P_{i,j} \frac{y_i}{(P\lambda^k)_i}. \tag{10.60}$$

The sequence $\{\lambda^k\}$ converges to a maximizer $\lambda_{ML}$ of the likelihood for any positive starting vector.

As we noted previously, before we can discuss the possible convergence of the sequence $\{\lambda^k\}$ of parameter vectors to a maximizer of the likelihood, it is necessary to have a notion of convergence in the parameter space. For the problem in this section, the parameter vectors $\lambda$ are non-negative.

Proof of convergence of the sequence $\{\lambda^k\}$ depends heavily on the following [30]:

$$KL(y, P\lambda^k) - KL(y, P\lambda^{k+1}) = KL(r(\lambda^k), r(\lambda^{k+1})) + KL(\lambda^{k+1}, \lambda^k), \quad (10.61)$$

and

$$KL(\lambda_{ML}, \lambda^k) - KL(\lambda_{ML}, \lambda^{k+1}) \geq KL(y, P\lambda^k) - KL(y, P\lambda_{ML}). \quad (10.62)$$

Any likelihood maximizer $\lambda_{ML}$ is also a non-negative minimizer of the KL distance $KL(y, P\lambda)$, so the EMML algorithm can be thought of as a method for finding a non-negative solution (or approximate solution) for a system $y = P\lambda$ of linear equations in which $y_i > 0$ and $P_{ij} \geq 0$ for all indices. This will be helpful when we consider mixture problems.

## 10.9 Finite Mixture Problems

Estimating the combining proportions in probabilistic mixture problems shows that there are meaningful examples of our acceptable-data model, and provides important applications of likelihood maximization.

### 10.9.1 Mixtures

We say that a random vector $V$ taking values in $R^M$ is a *finite mixture* (see Everett and Hand [81]; Redner and Walker [133]) if there are probability density functions or probabilities $f_j$ and numbers $\theta_j \geq 0$, for $j = 1, ..., J$, such that the probability density function or probability function for $V$ has the form

$$f_V(v|\theta) = \sum_{j=1}^{J} \theta_j f_j(v), \quad (10.63)$$

for some choice of the $\theta_j \geq 0$ with $\sum_{j=1}^{J} \theta_j = 1$.

### 10.9.2 The Likelihood Function

The data are $N$ realizations of the random vector $V$, denoted $v^n$, for $n = 1, ..., N$, and the given data is the vector $y = (v^1, ..., v^N)$. The column vector $\theta = (\theta_1, ..., \theta_J)^T$ is the generic parameter vector of mixture combining proportions. The likelihood function is

$$L_y(\theta) = \prod_{n=1}^{N} \Big(\theta_1 f_1(v^n) + ... + \theta_J f_J(v^n)\Big). \quad (10.64)$$

Then the log likelihood function is

$$LL_y(\theta) = \sum_{n=1}^{N} \log\Big(\theta_1 f_1(v^n) + ... + \theta_J f_J(v^n)\Big).$$

With $u$ the column vector with entries $u_i = 1/N$, and $P$ the matrix with entries $P_{nj} = f_j(v^n)$, we define

$$s_j = \sum_{n=1}^{N} P_{nj} = \sum_{n=1}^{N} f_j(v^n).$$

Maximizing $LL_y(\theta)$ is equivalent to minimizing

$$F(\theta) = KL(u, P\theta) + \sum_{j=1}^{J}(1 - s_j)\theta_j. \tag{10.65}$$

### 10.9.3 A Motivating Illustration

To motivate such mixture problems, we imagine that each data value is generated by first selecting one value of $j$, with probability $\theta_j$, and then selecting a realization of a random variable governed by $f_j(v)$. For example, there could be $J$ bowls of colored marbles, and we randomly select a bowl, and then randomly select a marble within the selected bowl. For each $n$ the number $v_n$ is the numerical code for the color of the $n$th marble drawn. In this illustration we are using a mixture of probability functions, but we could have used probability density functions.

### 10.9.4 The Acceptable Data

We approach the mixture problem by creating acceptable data. We imagine that we could have obtained $x_n = j_n$, for $n = 1, ..., N$, where the selection of $v^n$ is governed by the function $f_{j_n}(v)$. In the bowls example, $j_n$ is the number of the bowl from which the $n$th marble is drawn. The acceptable-data random vector is $X = (X_1, ..., X_N)$, where the $X_n$ are independent random variables taking values in the set $\{j = 1, ..., J\}$. The value $j_n$ is one realization of $X_n$. Since our objective is to estimate the true $\theta_j$, the values $v^n$ are now irrelevant. Our ML estimate of the true $\theta_j$ is simply the proportion of times $j = j_n$. Given a realization $x$ of $X$, the conditional pdf or pf of $Y$ does not involve the mixing proportions, so $X$ is acceptable. Notice also that it is not possible to calculate the entries of $y$ from those of $x$; the model $Y = h(X)$ does not hold.

### 10.9.5 The Mix-EM Algorithm

Using this acceptable data, we derive the EM algorithm, which we call the Mix-EM algorithm.

With $N_j$ denoting the number of times the value $j$ occurs as an entry of $x$, the likelihood function for $X$ is

$$L_x(\theta) = f_X(x|\theta) = \prod_{j=1}^{J} \theta_j^{N_j}, \tag{10.66}$$

and the log likelihood is

$$LL_x(\theta) = \log L_x(\theta) = \sum_{j=1}^{J} N_j \log \theta_j. \tag{10.67}$$

Then

$$E(\log L_x(\theta)|y, \theta^k) = \sum_{j=1}^{J} E(N_j|y, \theta^k) \log \theta_j. \tag{10.68}$$

To simplify the calculations in the E-step we rewrite $LL_x(\theta)$ as

$$LL_x(\theta) = \sum_{n=1}^{N} \sum_{j=1}^{J} X_{nj} \log \theta_j, \tag{10.69}$$

where $X_{nj} = 1$ if $j = j_n$ and zero otherwise. Then we have

$$E(X_{nj}|y, \theta^k) = \text{prob}\,(X_{nj} = 1|y, \theta^k) = \frac{\theta_j^k f_j(v^n)}{f(v^n|\theta^k)}. \tag{10.70}$$

The function $E(LL_x(\theta)|y, \theta^k)$ becomes

$$E(LL_x(\theta)|y, \theta^k) = \sum_{n=1}^{N} \sum_{j=1}^{J} \frac{\theta_j^k f_j(v^n)}{f(v^n|\theta^k)} \log \theta_j. \tag{10.71}$$

Maximizing with respect to $\theta$, we get the iterative step of the Mix-EM algorithm:

$$\theta_j^{k+1} = \frac{1}{N} \theta_j^k \sum_{n=1}^{N} \frac{f_j(v^n)}{f(v^n|\theta^k)}. \tag{10.72}$$

We know from our previous discussions that, since the preferred data $X$ is acceptable, likelihood is non-decreasing for this algorithm. We shall go further now, and show that the sequence of probability vectors $\{\theta^k\}$ converges to a maximizer of the likelihood.

### 10.9.6 Convergence of the Mix-EM Algorithm

As we noted earlier, maximizing the likelihood in the mixture case is equivalent to minimizing

$$F(\theta) = KL(u, P\theta) + \sum_{j=1}^{J}(1 - s_j)\theta_j,$$

over probability vectors $\theta$. It is easily shown that if $\hat{\theta}$ minimizes $F(\theta)$ over all non-negative vectors $\theta$, then $\hat{\theta}$ is a probability vector. Therefore, we can obtain the maximum likelihood estimate of $\theta$ by minimizing $F(\theta)$ over non-negative vectors $\theta$.

The following theorem is found in [39].

**Theorem 10.1** *Let $u$ be any positive vector, $P$ any non-negative matrix with $s_j > 0$ for each $j$, and*

$$F(\theta) = KL(u, P\theta) + \sum_{j=1}^{J}\beta_j KL(\gamma_j, \theta_j).$$

*If $s_j + \beta_j > 0$, $\alpha_j = s_j/(s_j + \beta_j)$, and $\beta_j\gamma_j \geq 0$, for all $j$, then the iterative sequence given by*

$$\theta_j^{k+1} = \alpha_j s_j^{-1}\theta_j^k\Big(\sum_{n=1}^{N} P_{n,j}\frac{u_n}{(P\theta^k)_n}\Big) + (1 - \alpha_j)\gamma_j \tag{10.73}$$

*converges to a non-negative minimizer of $F(\theta)$.*

With the choices $u_n = 1/N$, $\gamma_j = 0$, and $\beta_j = 1 - s_j$, the iteration in Equation (10.73) becomes that of the Mix-EM algorithm. Therefore, the sequence $\{\theta^k\}$ converges to the maximum likelihood estimate of the mixing proportions.

## 10.10 More on Convergence

There is a mistake in the proof of convergence given in Dempster, Laird, and Rubin (1977) [74]. Wu (1983) [153] and Boyles (1983) [14] attempted to repair the error, but also gave examples in which the EM algorithm failed to converge to a global maximizer of likelihood. In Chapter 3 of McLachlan and Krishnan (1997) [120] we find the basic theory of the EM algorithm, including available results on convergence and the rate of convergence. Because many authors rely on Equation (10.3), it is not clear that these results are valid in the generality in which they are presented.

There appears to be no single convergence theorem that is relied on universally; each application seems to require its own proof of convergence. When the use of the EM algorithm was suggested for SPECT and PET, it was necessary to prove convergence of the resulting iterative algorithm in Equation (10.58), as was eventually achieved in a sequence of papers (Shepp and Vardi (1982) [139], Lange and Carson (1984) [113], Vardi, Shepp and Kaufman (1985) [149], Lange, Bahn and Little (1987) [114], and Byrne (1993) [30]). When the EM algorithm was applied to list-mode data in SPECT and PET (Barrett, White, and Parra (1997) [6], and Huesman et al. (2000) [105], the resulting algorithm differed slightly from that in Equation (10.58) and a proof of convergence was provided in Byrne (2001) [39]. The convergence theorem in Byrne (2001) also establishes the convergence of the iteration in Equation (10.72) to the maximum-likelihood estimate of the mixing proportions.

## 10.11 Open Questions

As we have seen, the conventional formulation of the EM algorithm presents difficulties when probability density functions are involved. We have shown here that the use of acceptable preferred data can be helpful in resolving this issue, but other ways may also be useful.

Proving convergence of the sequence $\{\theta^k\}$ appears to involve the selection of an appropriate topology for the parameter space $\Theta$. While it is common to assume that $\Theta$ is a subset of Euclidean space and that the usual norm should be used to define distance, it may be helpful to tailor the metric to the nature of the parameters. In the case of Poisson sums, for example, the parameters are non-negative vectors and we found that the cross-entropy distance is more appropriate. Even so, additional assumptions appear necessary before convergence of the $\{\theta^k\}$ can be established. To simplify the analysis, it is often assumed that cluster points of the sequence lie in the interior of the set $\Theta$, which is not a realistic assumption in some applications.

It may be wise to consider, instead, convergence of the functions $f_Y(y|\theta^k)$, or maybe even to identify the parameters $\theta$ with the functions $f_Y(y|\theta)$. Proving convergence to $L_y(\theta_{ML})$ of the likelihood values $L_y(\theta^k)$ is also an option.

## 10.12 Conclusion

Difficulties with the conventional formulation of the EM algorithm in the continuous case of probability density functions (pdf) has prompted us to adopt a new definition, that of acceptable data. As we have shown, this

model can be helpful in generating EM algorithms in a variety of situations. For the discrete case of probability functions (pf), the conventional approach remains satisfactory. In both cases, the two steps of the EM algorithm can be viewed as alternating minimization of the Kullback-Leibler distance between two sets of parameterized pf or pdf, along the lines investigated by Csiszár and Tusnády [70]. In order to use the full power of their theory, however, we need one of the sets to be convex. This does occur in the important special case of sums of independent Poisson random variables, but is not generally the case.

# Chapter 11

# Kepler's Laws of Planetary Motion (Chapter 5,6)

## 11.1 Introduction

Kepler worked from 1601 to 1612 in Prague as the Imperial Mathematician. Taking over from Tycho Brahe, and using the tremendous amount of data gathered by Brahe from naked-eye astronomical observation, he formulated three laws governing planetary motion. Fortunately, among his tasks was the study of the planet Mars, whose orbit is quite unlike a circle. This forced Kepler to consider other possibilities and ultimately led to his discovery of elliptic orbits. These laws, which were the first "natural laws" in the modern sense, served to divorce astronomy from theology and marry it to physics. At last, the planets were viewed as material bodies, not unlike earth, floating freely in space and moved by physical forces acting on them. Although the second law preceded the first, Kepler's Laws are usually enumerated as follows:

- 1. the planets travel around the sun not in circles but in elliptical orbits, with the sun at one focal point;
- 2. a planet's speed is not uniform, but is such that the line segment from the sun to the planet sweeps out equal areas in equal time intervals; and, finally,
- 3. for all the planets, the time required for the planet to complete one orbit around the sun, divided by the 3/2 power of its average distance from the sun, is the same constant.

These laws, particularly the third one, provided strong evidence for Newton's law of universal gravitation. How Kepler discovered these laws without the aid of analytic geometry and differential calculus, with no notion of momentum, and only a vague conception of gravity, is a fascinating story, perhaps best told by Koestler in [108].

Around 1684, Newton was asked by Edmund Halley, of Halley's comet fame, what the path would be for a planet moving around the sun, if the force of gravity fell off as the square of the distance from the sun. Newton responded that it would be an ellipse. Kepler had already declared that planets moved along elliptical orbits with the sun at one focal point, but his findings were based on observation and imagination, not deduction from physical principles. Halley asked Newton to provide a proof. To supply such a proof, Newton needed to write a whole book, the *Principia*, published in 1687, in which he had to deal with such mathematically difficult questions as what the gravitational force is on a point when the attracting body is not just another point, but a sphere, like the sun.

With the help of vector calculus, a later invention, Kepler's laws can be derived as consequences of Newton's inverse square law for gravitational attraction.

## 11.2 Preliminaries

We consider a body with constant mass $m$ moving through three-dimensional space along a curve

$$\mathbf{r}(t) = (x(t), y(t), z(t)),$$

where $t$ is time and the sun is the origin. The velocity vector at time $t$ is then

$$\mathbf{v}(t) = \mathbf{r}'(t) = (x'(t), y'(t), z'(t)),$$

and the acceleration vector at time $t$ is

$$\mathbf{a}(t) = \mathbf{v}'(t) = \mathbf{r}''(t) = (x''(t), y''(t), z''(t)).$$

The *linear momentum* vector is

$$\mathbf{p}(t) = m\mathbf{v}(t).$$

One of the most basic laws of motion is that the vector $\mathbf{p}'(t) = m\mathbf{v}'(t) = m\mathbf{a}(t)$ is equal to the external force exerted on the body. When a body, or more precisely, the center of mass of the body, does not change location, all it can do is rotate. In order for a body to rotate about an axis a *torque* is required. Just as work equals force times distance moved, work done in rotating a body equals torque times angle through which it is rotated. Just as force is the time derivative of $\mathbf{p}(t)$, the linear momentum vector, we find that torque is the time derivative of something else, called the *angular momentum vector*.

## 11.3 Torque and Angular Momentum

Consider a body rotating around the origin in two-dimensional space, whose position at time $t$ is

$$\mathbf{r}(t) = (r\cos\theta(t), r\sin\theta(t)).$$

Then at time $t + \Delta t$ it is at

$$\mathbf{r}(t + \Delta t) = (r\cos(\theta(t) + \Delta\theta), r\sin(\theta(t) + \Delta\theta)).$$

Therefore, using trig identities, we find that the change in the $x$-coordinate is approximately

$$\Delta x = -r\Delta\theta\sin\theta(t) = -y(t)\Delta\theta,$$

and the change in the $y$-coordinate is approximately

$$\Delta y = r\Delta\theta\cos\theta(t) = x(t)\Delta\theta.$$

The infinitesimal work done by a force $\mathbf{F} = (F_x, F_y)$ in rotating the body through the angle $\Delta\theta$ is then approximately

$$\Delta W = F_x\Delta x + F_y\Delta y = (F_y x(t) - F_x y(t))\Delta\theta.$$

Since work is torque times angle, we define the torque to be

$$\tau = F_y x(t) - F_x y(t).$$

The entire motion is taking place in two dimensional space. Nevertheless, it is convenient to make use of the concept of cross product of three-dimensional vectors to represent the torque. When we rewrite

$$\mathbf{r}(t) = (x(t), y(t), 0),$$

and

$$\mathbf{F} = (F_x, F_y, 0),$$

we find that

$$\mathbf{r}(t) \times \mathbf{F} = (0, 0, F_y x(t) - F_x y(t)) = (0, 0, \tau) = \tau.$$

Now we use the fact that the force is the time derivative of the vector $\mathbf{p}(t)$ to write

$$\tau = (0, 0, \tau) = \mathbf{r}(t) \times \mathbf{p}'(t).$$

**Ex. 11.1** *Show that*

$$\mathbf{r}(t) \times \mathbf{p}'(t) = \frac{d}{dt}(\mathbf{r}(t) \times \mathbf{p}(t)). \tag{11.1}$$

By analogy with force as the time derivative of linear momentum, we define torque as the time derivative of angular momentum, which, from the calculations just performed, leads to the definition of the *angular momentum vector* as

$$\mathbf{L}(t) = \mathbf{r}(t) \times \mathbf{p}(t).$$

We need to say a word about the word "vector". In our example of rotation in two dimensions we introduced the third dimension as merely a notational convenience. It is convenient to be able to represent the torque as $\mathbf{L}'(t) = (0, 0, \tau)$, but when we casually call $L(t)$ the angular momentum vector, physicists would tell us that we haven't yet shown that angular momentum is a "vector" in the physicists' sense. Our example was too simple, they would point out. We had rotation about a single fixed axis that was conveniently chosen to be one of the coordinate axes in three-dimensional space. But what happens when the coordinate system changes?

Clearly, they would say, physical objects rotate and have angular momentum. The earth rotates around an axis, but this axis is not always the same axis; the axis wobbles. A well thrown football rotates around its longest axis, but this axis changes as the ball flies through the air. Can we still say that the angular momentum can be represented as

$$\mathbf{L}(t) = \mathbf{r}(t) \times \mathbf{p}(t)?$$

In other words, we need to know that the torque is still the time derivative of $\mathbf{L}(t)$, even as the coordinate system changes. In order for something to be a "vector"in the physicists' sense, it needs to behave properly as we switch coordinate systems, that is, it needs to *transform as a vector* [84]. In fact, all is well. This definition of $\mathbf{L}(t)$ holds for bodies moving along more general curves in three-dimensional space, and we can go on calling $\mathbf{L}(t)$ the angular momentum vector. Now we begin to exploit the special nature of the gravitational force.

## 11.4 Gravity is a Central Force

We are not interested here in arbitrary forces, but in the gravitational force that the sun exerts on the body, which has special properties that we shall exploit. In particular, this gravitational force is a *central force*.

**Definition 11.1** *We say that the force is a central force if*

$$\mathbf{F}(t) = h(t)\mathbf{r}(t),$$

*for each $t$, where $h(t)$ denotes a scalar function of $t$; that is, the force is central if it is proportional to $\mathbf{r}(t)$ at each $t$.*

**Proposition 11.1** *If $\mathbf{F}(t)$ is a central force, then $\mathbf{L}'(t) = \mathbf{0}$, for all $t$, so that $\mathbf{L} = \mathbf{L}(t)$ is a constant vector and $L = ||\mathbf{L}(t)|| = ||\mathbf{L}||$ is a constant scalar, for all $t$.*

**Proof:** From Equation (11.1) we have

$$\mathbf{L}'(t) = \mathbf{r}(t) \times \mathbf{p}'(t) = \mathbf{r}(t) \times \mathbf{F}(t) = h(t)\mathbf{r}(t) \times \mathbf{r}(t) = \mathbf{0}.$$

∎

We see then that the angular momentum vector $\mathbf{L}(t)$ is *conserved* when the force is central.

**Proposition 11.2** *If $\mathbf{L}'(t) = \mathbf{0}$, then the curve $\mathbf{r}(t)$ lies in a plane.*

**Proof:** We have

$$\mathbf{r}(t) \cdot \mathbf{L} = \mathbf{r}(t) \cdot \mathbf{L}(t) = \mathbf{r}(t) \cdot \Big(\mathbf{r}(t) \times \mathbf{p}(t)\Big),$$

which is the volume of the parallelepiped formed by the three vectors $\mathbf{r}(t)$, $\mathbf{r}(t)$ and $\mathbf{p}(t)$, which is obviously zero. Therefore, for every $t$, the vector $\mathbf{r}(t)$ is orthogonal to the constant vector $\mathbf{L}$. So, the curve lies in a plane with normal vector $\mathbf{L}$. ∎

## 11.5 The Second Law

We know now that, since the force is central, the curve described by $\mathbf{r}(t)$ lies in a plane. This allows us to use polar coordinate notation [144]. We write

$$\mathbf{r}(t) = \rho(t)(\cos\theta(t), \sin\theta(t)) = \rho(t)\mathbf{u}_r(t),$$

where $\rho(t)$ is the length of the vector $\mathbf{r}(t)$ and

$$\mathbf{u}_r(t) = \frac{\mathbf{r}(t)}{||\mathbf{r}(t)||} = (\cos\theta(t), \sin\theta(t))$$

is the unit vector in the direction of $\mathbf{r}(t)$. We also define

$$\mathbf{u}_\theta(t) = (-\sin\theta(t), \cos\theta(t)),$$

so that

$$\mathbf{u}_\theta(t) = \frac{d}{d\theta}\mathbf{u}_r(t),$$

and

$$\mathbf{u}_r(t) = -\frac{d}{d\theta}\mathbf{u}_\theta(t).$$

**Ex. 11.2** *Show that*

$$\mathbf{p}(t) = m\rho'(t)\mathbf{u}_r(t) + m\rho(t)\frac{d\theta}{dt}\mathbf{u}_\theta(t). \tag{11.2}$$

**Ex. 11.3** *View the vectors* $\mathbf{r}(t)$, $\mathbf{p}(t)$, $\mathbf{u}_r(t)$ *and* $\mathbf{u}_\theta(t)$ *as vectors in three-dimensional space, all with third component equal to zero. Show that*

$$\mathbf{u}_r(t) \times \mathbf{u}_\theta(t) = \mathbf{k} = (0, 0, 1),$$

*for all* $t$*. Use this and Equation (11.2) to show that*

$$\mathbf{L} = \mathbf{L}(t) = \Big(m\rho(t)^2\frac{d\theta}{dt}\Big)\mathbf{k},$$

*so that* $L = m\rho(t)^2\frac{d\theta}{dt}$*, the moment of inertia times the angular velocity, is constant.*

Let $t_0$ be some arbitrary time, and for any time $t \geq t_0$ let $A(t)$ be the area swept out by the planet in the time interval $[t_0, t]$. Then $A(t_2) - A(t_1)$ is the area swept out in the time interval $[t_1, t_2]$.

In the very short time interval $[t, t + \Delta t]$ the vector $\mathbf{r}(t)$ sweeps out a very small angle $\Delta\theta$, and the very small amount of area formed is then approximately

$$\Delta A = \frac{1}{2}\rho(t)^2\Delta\theta.$$

Dividing by $\Delta t$ and taking limits, as $\Delta t \to 0$, we get

$$\frac{dA}{dt} = \frac{1}{2}\rho(t)^2\frac{d\theta}{dt} = \frac{L}{2m}.$$

Therefore, the area swept out between times $t_1$ and $t_2$ is

$$A(t_2) - A(t_1) = \int_{t_1}^{t_2}\frac{dA}{dt}dt = \int_{t_1}^{t_2}\frac{L}{2m}dt = \frac{L(t_2 - t_1)}{2m}.$$

This is Kepler's Second Law.

## 11.6 The First Law

We saw previously that the angular momentum vector is conserved when the force is central. When Newton's inverse-square law holds, there is another conservation law; the *Runge-Lenz vector* is also conserved. We shall use this fact to derive the First Law.

Let $M$ denote the mass of the sun, and $G$ Newton's gravitational constant.

**Definition 11.2** *The force obeys Newton's inverse square law if*

$$\mathbf{F}(t) = h(t)\mathbf{r}(t) = -\frac{mMG}{\rho(t)^3}\mathbf{r}(t).$$

Then we can write

$$\mathbf{F}(t) = -\frac{mMG}{\rho(t)^2}\frac{\mathbf{r}(t)}{||\mathbf{r}(t)||} = -\frac{mMG}{\rho(t)^2}\mathbf{u}_r(t).$$

**Definition 11.3** *The* Runge-Lenz vector *is*

$$\mathbf{K}(t) = \mathbf{p}(t) \times \mathbf{L}(t) - k\mathbf{u}_r(t),$$

*where* $k = m^2MG$.

**Ex. 11.4** *Show that the velocity vectors* $\mathbf{r}'(t)$ *lie in the same plane as the curve* $\mathbf{r}(t)$.

**Ex. 11.5** *Use the rule*

$$\mathbf{A} \times (\mathbf{A} \times \mathbf{B}) = (\mathbf{A} \cdot \mathbf{B})\mathbf{A} - (\mathbf{A} \cdot \mathbf{A})\mathbf{B}$$

*to show that* $\mathbf{K}'(t) = \mathbf{0}$, *so that* $\mathbf{K} = \mathbf{K}(t)$ *is a constant vector and* $K = ||\mathbf{K}||$ *is a constant scalar.*

So the Runge-Lenz vector is conserved when the force obeys Newton's inverse square law.

**Ex. 11.6** *Use the rule in the previous exercise to show that the constant vector* $\mathbf{K}$ *also lies in the plane of the curve* $\mathbf{r}(t)$.

**Ex. 11.7** *Show that*

$$\mathbf{K} \cdot \mathbf{r}(t) = L^2 - k\rho(t).$$

It follows from this exercise that

$$L^2 - k\rho(t) = \mathbf{K} \cdot \mathbf{r}(t) = K\rho(t)\cos\alpha(t),$$

where $\alpha(t)$ is the angle between the vectors $\mathbf{K}$ and $\mathbf{r}(t)$. From this we get

$$\rho(t) = L^2/(k + K\cos\alpha(t)).$$

For $k > K$, this is the equation of an ellipse having eccentricity $e = K/k$. This is Kepler's First Law.

Kepler initially thought that the orbits were "egg-shaped" , but later came to realize that they were ellipses. Although Kepler did not have the

analytical geometry tools to help him, he was familiar with the mathematical development of ellipses in the *Conics*, the ancient book by Apollonius, written in Greek in Alexandria about 200 BC. Conics, or conic sections, are the terms used to describe the two-dimensional curves, such as ellipses, parabolas and hyperbolas, formed when a plane intersects an infinite double cone (think "hour-glass").

Apollonius was interested in astronomy and Ptolemy was certainly aware of the work of Apollonius, but it took Kepler to overcome the bias toward circular motion and introduce conic sections into astronomy. As related by Bochner [13], there is a bit of mystery concerning Kepler's use of the *Conics*. He shows that he is familiar with a part of the *Conics* that existed only in Arabic until translated into Latin in 1661, well after his time. How he gained that familiarity is the mystery.

## 11.7 The Third Law

As the planet moves around its orbit, the closest distance to the sun is

$$\rho_{\min} = L^2/(k+K),$$

and the farthest distance is

$$\rho_{\max} = L^2/(k-K).$$

The average of these two is

$$a = \frac{1}{2}\Big(\rho_{\min} + \rho_{\max}\Big) = 2kL^2/(k^2-K^2);$$

this is the semi-major axis of the ellipse. The semi-minor axis has length $b$, where

$$b^2 = a^2(1-e^2).$$

Therefore,

$$b = \frac{L\sqrt{a}}{\sqrt{k}}.$$

The area of this ellipse is $\pi ab$. But we know from the first law that the area of the ellipse is $\frac{L}{2m}$ times the time $T$ required to complete a full orbit. Equating the two expressions for the area, we get

$$T^2 = \frac{2\pi}{MG}a^3.$$

This is the third law.

The first two laws deal with the behavior of one planet; the third law is different. The third law describes behavior that is common to all the planets in the solar system, thereby suggesting a universality to the force of gravity.

## 11.8 From Kepler to Newton

Our goal, up to now, has been to show how Kepler's three laws can be derived from Newton's inverse-square law, which, of course, is not how Kepler obtained the laws. Kepler arrived at his laws empirically, by studying the astronomical data. Newton was aware of Kepler's laws and they influenced his work on universal gravitation. When asked what would explain Kepler's elliptical orbits, Newton replied that he had calculated that an inverse-square law would do it. Newton found that the force required to cause the moon to deviate from a tangent line was approximately that given by an inverse-square fall-off in gravity.

It is interesting to ask if the inverse-square law can be derived from Kepler's three laws; the answer is yes, as we shall see in this section. What follows is taken from [94].

We found previously that

$$\frac{dA}{dt} = \frac{1}{2}\rho(t)^2\frac{d\theta}{dt} = \frac{L}{2m} = c. \tag{11.3}$$

Differentiating with respect to $t$, we get

$$\rho(t)\rho'(t)\frac{d\theta}{dt} + \frac{1}{2}\rho(t)^2\frac{d^2\theta}{dt^2} = 0, \tag{11.4}$$

so that

$$2\rho'(t)\frac{d\theta}{dt} + \rho(t)\frac{d^2\theta}{dt^2} = 0. \tag{11.5}$$

From this, we shall prove that the force is central, directed towards the sun.

As we did earlier, we write the position vector $\mathbf{r}(t)$ as

$$\mathbf{r}(t) = \rho(t)\mathbf{u}_r(t),$$

so, suppressing the dependence on the time $t$, and using the identities

$$\frac{d\mathbf{u}_r}{dt} = \mathbf{u}_\theta\frac{d\theta}{dt},$$

and

$$\frac{d\mathbf{u}_\theta}{dt} = -\mathbf{u}_r\frac{d\rho}{dt},$$

we write the velocity vector as

$$\mathbf{v} = \frac{d\mathbf{r}}{dt} = \frac{d\rho}{dt}\mathbf{u}_r + \rho\frac{d\mathbf{u}_r}{dt} = \frac{d\rho}{dt}\mathbf{u}_r + \rho\frac{d\mathbf{u}_r}{d\theta}\frac{d\theta}{dt} = \frac{d\rho}{dt}\mathbf{u}_r + \rho\frac{d\theta}{dt}\mathbf{u}_\theta,$$

and the acceleration vector as

$$\mathbf{a} = \frac{d^2\rho}{dt^2}\mathbf{u}_r + \frac{d\rho}{dt}\frac{d\mathbf{u}_r}{dt} + \frac{d\rho}{dt}\frac{d\theta}{dt}\mathbf{u}_\theta + \rho\frac{d^2\theta}{dt^2}\mathbf{u}_\theta + \rho\frac{d\theta}{dt}\frac{d\mathbf{u}_\theta}{dt}$$

$$= \frac{d^2\rho}{dt^2}\mathbf{u}_r + \frac{d\rho}{dt}\frac{d\theta}{dt}\mathbf{u}_\theta + \frac{d\rho}{dt}\frac{d\theta}{dt}\mathbf{u}_\theta + \rho\frac{d^2\theta}{dt^2}\mathbf{u}_\theta - \rho\frac{d\theta}{dt}\frac{d\theta}{dt}\mathbf{u}_r.$$

Therefore, we have

$$\mathbf{a} = \Big(\frac{d^2\rho}{dt^2} - \rho(\frac{d\theta}{dt})^2\Big)\mathbf{u}_r + \Big(2\frac{d\rho}{dt}\frac{d\theta}{dt} + \rho\frac{d^2\theta}{dt^2}\Big)\mathbf{u}_\theta.$$

Using Equation (11.4), this reduces to

$$\mathbf{a} = \Big(\frac{d^2\rho}{dt^2} - \rho(\frac{d\theta}{dt})^2\Big)\mathbf{u}_r, \tag{11.6}$$

which tells us that the acceleration, and therefore the force, is directed along the line joining the planet to the sun; it is a central force.

**Ex. 11.8** *Prove the following two identities:*

$$\frac{d\rho}{dt} = \frac{d\rho}{d\theta}\frac{d\theta}{dt} = \frac{2c}{\rho^2}\frac{d\rho}{dt} \tag{11.7}$$

*and*

$$\frac{d^2\rho}{dt^2} = \frac{4c^2}{\rho^4}\frac{d^2\rho}{d\theta^2} - \frac{8c^2}{\rho^5}\Big(\frac{d\rho}{d\theta}\Big)^2. \tag{11.8}$$

Therefore, we can write the acceleration vector as

$$\mathbf{a} = \left(\frac{4c^2}{\rho^4}\frac{d^2\rho}{d\theta^2} - \frac{8c^2}{\rho^5}\Big(\frac{d\rho}{d\theta}\Big)^2 - \frac{4c^2}{\rho^3}\right)\mathbf{u}_r.$$

To simplify, we substitute $u = \rho^{-1}$.

**Ex. 11.9** *Prove that the acceleration vector can be written as*

$$\mathbf{a} = \left(4c^2u^2\left(-\frac{1}{u^2}\frac{d^2u}{d\theta^2} + \frac{2}{u^3}\Big(\frac{du}{d\theta}\Big)^2\right) - 8c^2u^5\Big(-\frac{1}{u^2}\frac{du}{d\theta}\Big)^2 - 4c^2u^3\right)\mathbf{u}_r,$$

*so that*

$$\mathbf{a} = -4c^2u^2\Big(\frac{d^2u}{d\theta^2} + u\Big)\mathbf{u}_r. \tag{11.9}$$

Kepler's First Law tells us that

$$\rho(t) = \frac{L^2}{k + K\cos\alpha(t)} = \frac{a(1-e^2)}{1 + e\cos\alpha(t)},$$

where $e = K/k$ and $a$ is the semi-major axis. Therefore,

$$u = \frac{1 + e\cos\alpha(t)}{a(1-e^2)}.$$

Using Equation (11.9), we can write the acceleration as

$$\mathbf{a} = -\frac{4c^2}{a(1-e^2)}u^2\mathbf{u}_r = -\frac{4c^2}{a(1-e^2)}r^{-2}\mathbf{u}_r,$$

which tells us that the force obeys an inverse-square law. We still must show that this same law applies to each of the planets, that is, that the constant $\frac{c^2}{a(1-e^2)}$ does not depend on the particular planet.

**Ex. 11.10** *Show that*

$$\frac{c^2}{a(1-e^2)} = \frac{\pi^2 a^3}{T^2},$$

*which is independent of the particular planet, according to Kepler's Third Law.*

## 11.9 Newton's Own Proof of the Second Law

Although Newton invented calculus, he relied on geometry for many of his mathematical arguments. A good example is his proof of Kepler's Second Law.

He begins by imagining the planet at the point 0 in Figure 11.1. If there were no force coming from the sun, then, by the principle of inertia, the planet would continue in a straight line, with constant speed. The distance $\Delta$ from the point 0 to the point 1 is the same as the distance from 1 to 2 and the same as the distance from 2 to 3. The areas of the three triangles formed by the sun and the points 0 and 1, the sun and the points 1 and 2, and the sun and the points 2 and 3 are all equal, since they all equal half of the base $\Delta$ times the height $H$. Therefore, in the absence of a force from the sun, the planet sweeps out equal areas in equal times. Now what happens when there is a force from the sun?

Newton now assumes that $\Delta$ is very small, and that during the short time it would have taken for the planet to move from 1 to 3 there is a force on the planet, directed toward the sun. Because of the small size of $\Delta$, he safely assumes that the direction of this force is unchanged and is directed along the line from 2, the midpoint of 1 and 3, to the sun. The effect of

such a force is to pull the planet away from 3, along the line from 3 to 4. The areas of the two triangles formed by the sun and the points 2 and 3 and the sun and the points 2 and 4 are both equal to half of the distance from the sun to 2, times the distance from 2 to $B$. So we still have equal areas in equal times.

We can corroborate Newton's approximations using vector calculus. Consider the planet at 2 at time $t = 0$. Suppose that the acceleration is $\mathbf{a}(t) = (b, c)$, where $(b, c)$ is a vector parallel to the line segment from the sun to 2. Then the velocity vector is $\mathbf{v}(t) = t(b, c) + (0, \Delta)$, where, for simplicity, we assume that, in the absence of the force from the sun, the planet travels at a speed of $\Delta$ units per second. The position vector is then

$$\mathbf{r}(t) = \frac{1}{2}t^2(b, c) + t(0, \Delta) + \mathbf{r}(0).$$

At time $t = 1$, instead of the planet being at 3, it is now at

$$\mathbf{r}(1) = \frac{1}{2}(b, c) + (0, \Delta) + \mathbf{r}(0).$$

Since the point 3 corresponds to the position $(0, \Delta) + \mathbf{r}(0)$, we see that the point 4 lies along the line from 3 parallel to the vector $(b, c)$.

## 11.10 Armchair Physics

Mathematicians tend to ignore things like units, when they do calculus problems. Physicists know that you can often learn a lot just by paying attention to the units involved, or by asking questions like what happens to velocity when length is converted from feet to inches and time from minutes to seconds. This is sometimes called "armchair physics" . To illustrate, we apply this approach to Kepler's Third Law.

### 11.10.1 Rescaling

Suppose that the spatial variables $(x, y, z)$ are replaced by $(\alpha x, \alpha y, \alpha z)$ and time changed from $t$ to $\beta t$. Then velocity, since it is distance divided by time, is changed from $v$ to $\alpha\beta^{-1}v$. Velocity squared, and therefore kinetic and potential energies, are changed by a factor of $\alpha^2\beta^{-2}$.

### 11.10.2 Gravitational Potential

The gravitational potential function $\phi(x, y, z)$ associated with the gravitational field due to the sun is given by

$$\phi(x, y, z) = \frac{-C}{\sqrt{x^2 + y^2 + z^2}}, \tag{11.10}$$

where $C > 0$ is some constant and we assume that the sun is at the origin. The gradient of $\phi(x, y, z)$ is

$$\nabla\phi(x,y,z) = \Big(\frac{-C}{x^2+y^2+z^2}\Big)\Big(\frac{x}{\sqrt{x^2+y^2+z^2}}, \frac{y}{\sqrt{x^2+y^2+z^2}}, \frac{z}{\sqrt{x^2+y^2+z^2}}\Big).$$

The gravitational force on a massive object at point $(x, y, z)$ is therefore a vector of magnitude $\frac{C}{x^2+y^2+z^2}$, directed from $(x, y, z)$ toward $(0, 0, 0)$, which says that the force is central and falls off as the reciprocal of the distance squared.

The potential function $\phi(x, y, z)$ is $(-1)$-homogeneous, meaning that when we replace $x$ with $\alpha x$, $y$ with $\alpha y$, and $z$ with $\alpha z$, the new potential is the old one times $\alpha^{-1}$.

We also know, though, that when we rescale the space variables by $\alpha$ and time by $\beta$ the potential energy is multiplied by a factor of $\alpha^2\beta^{-2}$. It follows that

$$\alpha^{-1} = \alpha^2\beta^{-2},$$

so that

$$\beta^2 = \alpha^3. \tag{11.11}$$

Suppose that we have two planets, $P_1$ and $P_2$, orbiting the sun in circular orbits, with the length of the the orbit of $P_2$ equal to $\alpha$ times that of $P_1$. We can view the orbital data from $P_2$ as that from $P_1$, after a rescaling of the spatial variables by $\alpha$. According to Equation (11.11), the orbital time of $P_2$ is then that of $P_1$ multiplied by $\beta = \alpha^{3/2}$. This is Kepler's Third Law.

Kepler took several decades to arrive at his third law, which he obtained not from basic physical principles, but from analysis of observational data. Could he have saved himself much time and effort if he had stayed in his armchair and considered rescaling, as we have just done? No. The importance of Kepler's Third Law lies in its universality, the fact that it applies not just to one planet but to all. We have implicitly assumed universality by postulating a potential function that governs the gravitational field from the sun.

### 11.10.3 Gravity on Earth

We turn now to the gravitational pull of the earth on an object near its surface. We have just seen that the potential function is proportional to the reciprocal of the distance from the center of the earth to the object. Let the radius of the earth be $R$ and let the object be at a height $h$ above the surface of the earth. Then the potential is

$$\phi(R+h) = \frac{-B}{R+h},$$

for some constant $B$. The potential at the surface of the earth is

$$\phi(R) = \frac{-B}{R}.$$

The potential difference between the object at height $h$ and the surface of the earth is then

$$PD(h) = \frac{B}{R} - \frac{B}{R+h} = B\Big(\frac{1}{R} - \frac{1}{R+h}\Big) = B\Big(\frac{R+h-R}{R(R+h)}\Big).$$

If $h$ is very small relative to $R$, then we can say that

$$PD(h) = \frac{B}{R^2}h,$$

so is linear in $h$. The potential difference is therefore 1-homogeneous; if we rescale the spatial variables by $\alpha$ the potential difference is also rescaled by $\alpha$. But, as we saw previously, the potential difference is also rescaled by $\alpha^2\beta^{-2}$. Therefore,

$$\alpha = \alpha^2\beta^{-2},$$

or

$$\beta = \alpha^{1/2}.$$

This makes sense. Consider a ball dropped from a tall building. In order to double the time of fall (multiply $t$ by $\beta = 2$) we must quadruple the height from which it is dropped (multiply $h$ by $\alpha = \beta^2 = 4$).

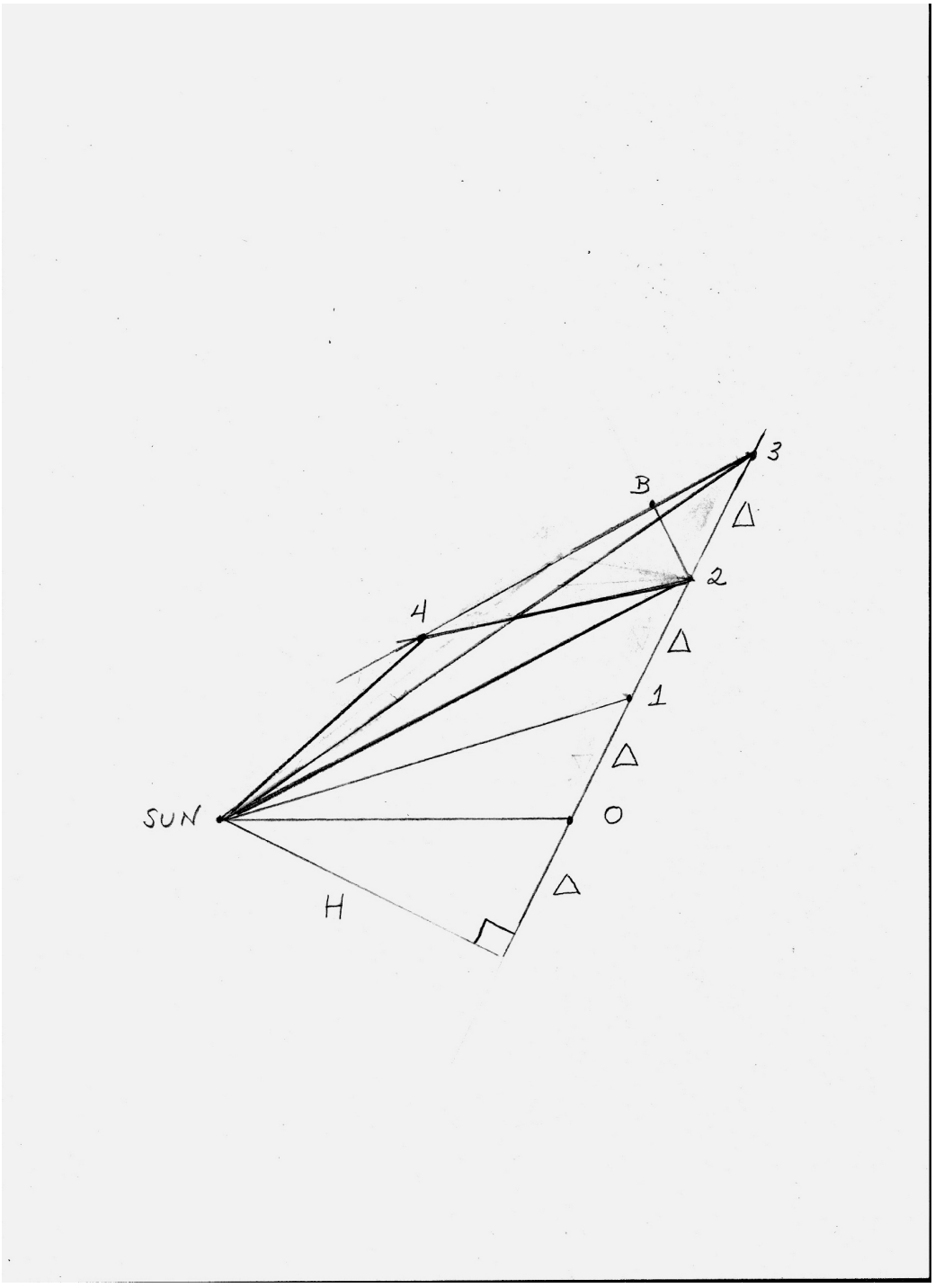


Figure 11.1: Newton's Own Diagram.

# Chapter 12

# A Brief History of Electromagnetism (Chapter 5,6)

## 12.1 Overview

Understanding the connections between magnetism and electricity and exploiting that understanding for technological innovation dominated science in the nineteenth century, and yet no one saw it coming. In the index to Butterfield's classic history of the scientific revolution [22], which he locates roughly from 1300 to 1800, the word "electricity" does not appear.

Electricity, as we now call it, was not completely unknown, of course. In the late sixteenth century, Gilbert, famous for his studies of magnetism, discovered that certain materials, mainly crystals, could be made attractive by rubbing them with a cloth. He called these materials *electrics*. Among Gilbert's accomplishments was his overturning of the conventional wisdom about magnets, when he showed, experimentally, that magnets *could* still attract nails after being rubbed with garlic. Sometime after Gilbert, electrostatic repulsion and induction were discovered, making the analogy with magnetism obvious. However, until some way was found to study electricity in the laboratory, the mysteries of electricity would remain hidden and its importance unappreciated.

Nobody in 1800 could have imagined that, within a hundred years or so, people would live in cities illuminated by electric light, work with machinery driven by electricity, in factories cooled by electric-powered refrigeration, and go home to listen to a radio and talk to neighbors on a telephone. How we got there is the subject of this note.

## 12.2 "What's Past is Prologue"

The history of science is not simply important for its own sake, but as a bridge connecting the arts with the sciences. When we study the history of science, we begin to see science as an integral part of the broader quest by human beings to understand themselves and their world. Progress in science comes not only from finding answers to questions, but from learning to ask better questions. The questions we are able to ask, indeed the observations we are able to make, are conditioned by our society, our history, and our intellectual outlook. Science does not exist in a vacuum. As Shakespeare's line, carved into the wall of the National Archives building in Washington, D.C., suggests, the past sets the stage for what comes next, indeed, for what can come next.

## 12.3 Are We There Yet?

We should be a little careful when we talk about progress, either within science or more generally. Reasonable people can argue about whether or not the development of atomic weapons ought to be called progress. Einstein and others warned, at the beginning of the atomic age, that the emotional and psychological development of human beings had not kept pace with technological development, that we did not have the capacity to control our technology. It does seem that we have a difficult time concerning ourselves, as a society, with problems that will become more serious in the future, preferring instead the motto "I won't be there. You won't be there."

We can certainly agree, though, that science, overall, has led us to a better, even if not complete, understanding of ourselves and our world and to the technology that is capable of providing decent life and health to far more people than in the past. These successes have given science and scientists a certain amount of political power that is not universally welcomed, however. Recent attempts to challenge the status of science within the community, most notably in the debate over creation "science" and evolution, have really been attempts to lessen the political power of science, not debates within science itself; the decades long attacks on science by the cigarette industry and efforts to weaken the EPA show clearly that it is not only some religious groups that want the political influence of science diminished.

Many of the issues our society will have to deal with in the near future, including nuclear power, terrorism, genetic engineering, energy, climate change, control of technology, space travel, and so on, involve science and demand a more sophisticated understanding of science on the part of the general public. The recent book *Physics for Future Presidents: the Science Behind the Headlines* [127] discusses many of these topics, supposedly as

an attempt by the author to educate presidents-to-be, who will be called on to make decisions, to initiate legislation, and to guide the public debate concerning these issues.

History reminds us that progress need not be permanent. The technological expertise and artistic heights achieved by the Romans, even the mathematical sophistication of Archimedes, were essentially lost, at least in the west, for fifteen hundred years.

History also teaches us how unpredictable the future can be, which is, in fact, the underlying theme of this chapter. No one in 1800 could have imagined the electrification that transformed society over the nineteenth century, just as no one in 1900 could have imagined Hiroshima and Nagasaki, only a few decades in the future, let alone the world of today.

## 12.4 Why Do Things Move?

In his famous "The Origins of Modern Science" [22] Butterfield singles out the problem of motion as the most significant intellectual hurdle the human mind has confronted and overcome in the last fifteen hundred years. The ancients had theories of motion, but for Aristotle, as a scientist perhaps more of a biologist than a physicist, motion as change in location was insignificant compared to motion as qualitative change, as, say, when an acorn grows into a tree. The change experienced by the acorn is clearly oriented toward a goal, to make a tree. By focusing on qualitative change, Aristotle placed too much emphasis on the importance of a goal. His idea that even physical motion was change toward a goal, that objects had a "natural" place to which they "sought" to return, infected science for almost two thousand years.

We must not be too quick to dismiss Aristotle's view, however. General relativity asserts that space-time is curved and that clocks slow down where gravity is stronger. Indeed, a clock on the top of the Empire State Building runs slightly faster than one at street level. As Brian Greene puts it,

*Right now, according to these ideas, you are anchored to the floor because your body is trying to slide down an indentation in space (really, spacetime) caused by the earth. In a sense, all objects "want" to age as slowly as possible* [96].

The one instance of motion as change in location whose importance the ancients appreciated was the motion of the heavens. Aristotle had his theories of the heavens and Ptolemy his astronomical system of an earth-centered universe. Because the objects in the heavens, the moon, the planets and the stars, certainly appear to move rapidly, they must be made of an unearthly material, the *quintessence*. So things stood until the middle ages. In the fourteenth century the French theologian Nicole Oresme considered the possibility that the earth rotated daily around its

own axis [117]. This hypothesis certainly simplified things considerably, and removed the need for the heavens to spin around the earth daily at enormous speeds. But even Oresme himself was hesitant to push this idea, since it conflicted with scripture.

Gradually, natural philosophers, the term used to describe scientists prior to the nineteenth century, began to take a more serious interest in motion as change in location, due, in part, to their growing interest in military matters and the motion of cannon balls. Now, motion on earth and motion of the heavenly bodies came to be studied by some of the same people, such as Galileo, and this set the stage for the unified theory of motion due to gravity that would come later, with Newton.

Copernicus' theory of a sun-centered astronomical system, Tycho Brahe's naked-eye observations of the heavens, Kepler's systematizing of planetary motion, the invention of the telescope and its use by Galileo to observe the pock-marked moon and the mini-planetary system of Jupiter, Galileo's study of balls rolling down inclined planes, and finally Newton's Law of Universal Gravitation marked a century of tremendous progress in the study of motion and put mechanics at the top of the list of scientific paradigms for the next century. Most of the theoretical developments of the eighteenth century involved the expansion of Newton's mechanics to ever more complex systems, so that, by the end of that century, celestial mechanics and potential theory were well developed mathematical subjects.

As we shall see, the early development of the field we now call electromagnetism involved little mathematics. As the subject evolved, the mathematics of potential theory, borrowed from the study of gravitation and celestial mechanics, was combined with the newly discovered vector calculus and the mathematical treatment of heat propagation to give the theoretical formulation of electromagnetism familiar to us today.

## 12.5 Go Fly a Kite

The ancients knew about magnets and used them as compasses. Static electricity was easily observed and thought to be similar to magnetism. As had been known for centuries, static electricity exhibited both attraction and repulsion. For that reason, it was argued that there were two distinct types of electricity. Benjamin Franklin opposed this idea, insisting instead on two types of charge, positive and negative. Some progress was made in capturing electricity for study with the invention of the *Leyden jar*, a device for storing relatively large electrostatic charge (and giving rather large shocks). The discharge from the Leyden jar reminded Franklin of lightning and prompted him and others to fly kites in thunderstorms and to discover that lightning would charge a Leyden jar; lightning was electricity. These experiments led to his invention of the lightning rod, a conducting

device attached to houses to direct lightning strikes down to the ground.

The obvious analogies with magnetism had been noticed by Gilbert and others in the late sixteenth century, and near the end of the eighteenth century Coulomb found that both magnetic and electrical attraction fell off as the square of the distance, as did gravity, according to Newton. Indeed, the physical connection between magnetism and gravity seemed more plausible than one between magnetism and electricity, and more worth studying. But things were about to change.

## 12.6 Bring in the Frogs

In 1791 Galvani observed that a twitching of the muscles of a dead frog he was dissecting seemed to be caused by sparks from a nearby discharge of a Leyden jar. He noticed that the sparks need not actually touch the muscles, provided a metal scalpel touched the muscles at the time of discharge. He also saw twitching muscles when the frog was suspended by brass hooks on an iron railing in a thunderstorm. Eventually, he realized that the Leyden jar and thunderstorm played no essential roles; two scalpels of different metals touching the muscles were sufficient to produce the twitching. Galvani concluded that the electricity was in the muscles; it was *animal electricity*.

## 12.7 Lose the Frogs

In 1800 Volta discovered that electricity could be produced by two dissimilar metals, copper and zinc, say, in salt water; no animal electricity here, and no further need for the frogs. He had discovered the *battery* and introduced *electrodynamics*. Only six weeks after Volta's initial report, Nicholson and Carlisle discovered *electrolysis*, the loosening up and separating of distinct atoms in molecules, such as the hydrogen and oxygen atoms in water.

The fact that chemical reactions produced electric currents suggested the reverse, that electrical currents could stimulate chemical reactions; this is *electrochemistry*, which led to the discovery and isolation of many new elements in the decades that followed. In 1807 Humphry Davy isolated some active metals from their liquid compounds and became the first to form sodium, potassium, calcium, strontium, barium, and magnesium.

In 1821 Seebeck found that the electric current would continue as long as the temperatures of the two metals were kept different; this is *thermoelectricity* and provides the basis for the *thermocouple*, which could then be used as a thermometer.

## 12.8 Bring in the Magnets

In 1819 Oersted placed a current-carrying wire over a compass, not expecting anything in particular to happen. The needle turned violently perpendicular to the axis of the wire. When Oersted reversed the direction of the current, the needle jerked around 180 degrees. This meant that magnetism and electricity were not just analogous, but intimately related; *electromagnetism* was born. Soon after, Arago demonstrated that a wire carrying an electric current behaved like a magnet. Ampere, in 1820, confirmed that a wire carrying a current *was* a magnet by demonstrating attraction and repulsion between two separate current-carrying wires. He also experimented with wires in various configurations and related the strength of the magnetic force to the strength of the current in the wire. This connection between electric current and magnetism led fairly soon after to the telegraph, and later in the century, to the telephone.

## 12.9 Enter Faraday

So electric currents can produce magnetism. But can magnets produce electric currents? Can the relationship be reversed? In 1831, Michael Faraday tried to see if a current would be produced in a wire if it was placed in a magnetic field created by another current-carrying wire. The experiment failed, sort of. When the current was turned on in the second wire, generating the magnetic field, the first wire experienced a brief current, but then nothing; when the current was turned off, again a brief current in the first wire. Faraday, an experimental genius who, as a young man, had been an assistant to Davy, and later the inventor of refrigeration, made the right conjecture that it is not the mere presence of the magnetic field that causes a current, but changes in that magnetic field. He confirmed this conjecture by showing that a current would flow through a coiled wire when a magnetized rod was moved in and out of the coil; he (and, independently, Henry in the USA) had invented *electromagnetic induction* and the *electric generator* and, like Columbus, had discovered a new world.

## 12.10 Do The Math

Mathematics has yet to appear in our brief history of electromagnetism, but that was about to change. Although Faraday, often described as being innocent of mathematics, developed his concept of *lines of force* in what we would view as an unsophisticated manner, he was a great scientist and his intuition would prove to be remarkably accurate.

In the summer of 1831, the same summer in which the forty-year old Faraday first observed the phenomenon of electromagnetic induction, the

creation of an electric current by a changing magnetic field, James Clerk Maxwell was born in Edinburgh, Scotland.

Maxwell's first paper on electromagnetism, "On Faraday's Lines of Force" , appeared in 1855, when he was about 25 years old. The paper involved a mathematical development of the results of Faraday and others and established the mathematical methods Maxwell would use later in his more famous work "On Physical Lines of Force" .

Although Maxwell did not have available all of the compact vector notation we have today, his work was mathematically difficult. The following is an excerpt from a letter Faraday himself sent to Maxwell concerning this point.

*There is one thing I would be glad to ask you. When a mathematician engaged in investigating physical actions and results has arrived at his conclusions, may they not be expressed in common language as fully, clearly and definitely as in mathematical formulae? If so, would it not be a great boon to such as I to express them so? - translating them out of their hieroglyphics, that we may work upon them by experiment.* Hasn't every beginning student of vector calculus and electromagnetism wished that Maxwell and his followers had heeded Faraday's pleas?

Maxwell reasoned that, since an electric current sets up a magnetic field, and a changing magnetic field creates an electrical field, there should be what we now call *electromagnetic waves*, as these two types of fields leapfrog across (empty?) space. These waves would obey partial differential equations, called *Maxwell's equations*, although their familiar form came later and is due to Heaviside [92]. Analyzing the mathematical properties of the resulting wave equations, Maxwell discovered that the propagation speed of these waves was the same as that of light, leading to the conclusion that light itself is an electromagnetic phenomenon, distinguished from other electromagnetic radiation only by its frequency. That light also exhibits behavior more particle-like than wave-like is part of the story of the science of the next century.

Maxwell predicted that electromagnetic radiation could exist at various frequencies, not only those associated with visible light. Infrared and ultraviolet radiation had been known since early in the century, and perhaps they too were part of a *spectrum* of electromagnetic radiation. After Maxwell's death from cancer at forty-eight, Hertz demonstrated, in 1888, the possibility of electromagnetic radiation at very low frequencies, *radio waves*. In 1895 Röntgen discovered electromagnetic waves at the high-frequency end of the spectrum, the so-called *x-rays*.

## 12.11 Just Dot the i's and Cross the t's?

By the end of the nineteenth century, some scientists felt that all that was left to do in physics was to dot the i's and cross the t's. However, other scientists saw paradoxes and worried that there were problems yet to be solved; how serious these might turn out to be was not always clear.

Maxwell himself had noted, about 1869, that his work on the specific heats of gases revealed conflicts between rigorous theory and experimental findings that he was unable to explain; it seemed that internal vibration of atoms was being "frozen out"at sufficiently low temperatures, something for which classical physics could not account. His was probably the first suggestion that classical physics could be "wrong". There were also the mysteries, observed by Newton, associated with the partial reflection of light by thick glass. Advances in geology and biology had suggested strongly that the earth and the sun were much older than previously thought, which was not possible, according to the physics of the day; unless a new form of energy was operating, the sun would have burned out a long time ago.

Newton thought that light was a stream of particles. Others at the time, notably Robert Hooke and Christiaan Huygens, felt that light was a wave phenomenon. Both sides were hindered by a lack of a proper scientific vocabulary to express their views.

Around 1800 Young demonstrated that a beam of light displayed interference effects similar to water waves. Eventually, his work convinced people that Newton had been wrong on this point and most accepted that light is a wave phenomenon. Faraday, Maxwell, Hertz and others further developed the wave theory of light and related light to other forms of electromagnetic radiation. Ironically, it was Hertz, in 1887, who discovered the *photo-electric effect*, later given by Einstein as confirming evidence that light has a particle nature. When light strikes a metal, it can cause the metal to release an electrically charged particle, an electron. If light were simply a wave, there would not be enough energy in the small part of the wave that hits the metal to displace the electron; in 1905 Einstein will argue that light is *quantized*, that is, it consists of individual bundles or particles, later called *photons*, each with enough energy to cause the electron to be released.

It was recognized that there were other problems with the wave theory of light. All known waves required a medium in which to propagate. Sound cannot propagate in a vacuum; it needs air or water or something. The sound waves are actually compressions and rarefactions of the medium, and how fast the waves propagate depends on how fast the material in the medium can perform these movements; sound travels faster in water than in air, for example.

Light travels extremely fast, but does not propagate instantaneously, as Olaus Roemer first demonstrated around 1700. He observed that the

eclipses of the moons of Jupiter appeared to happen sooner when Jupiter was moving closer to Earth, and later when it was moving away. He reasoned, correctly, that the light takes a finite amount of time to travel from the moons to Earth, and when Jupiter is moving away the distance is growing longer.

If light travels through a medium, which scientists called the *ether*, then the ether must be a very strange substance indeed. The material that makes up the ether must be able to compress and expand very quickly. Light comes to us from great distances so the ether must extend throughout all of space. The earth moves around the sun, and therefore through this ether, at a very great speed, and yet there are no friction effects, while very much slower winds produce a great deal of weathering. Light can also be polarized, so the medium must be capable of supporting transverse waves, not just longitudinal waves, as in acoustics. To top it all off, the Michelson-Morley experiment, performed in Cleveland in 1887, failed to detect the presence of the ether.

## 12.12 Seeing is Believing

If radio waves could travel around the earth through an invisible ether, and if hypnotists can *mesmerize* their subjects, why can't human beings communicate with each other and with the dead, telepathically? Why should atoms exist when we cannot see them, while ghosts must not, even when, as some claimed, they have shown up in photographs? When is seeing believing?

In the late 1800's the experimental physicist William Crooke claimed to have discovered *radiant matter* [82]. When he passed an electric current through a glass tube filled with a low-pressure gas, a small object within the tube could be made to move from one end to the other, driven, so Crooke claimed, by radiant particles of matter, later called *cathode rays*, streaming from one end of the tube to the other. Crooke then went on, without much success, to find material explanation for some of the alleged effects of spiritualism. He felt that it ought to be possible for humans to receive transmissions in much the same way as a radio receives signals. It was a time of considerable uncertainty, and it was not clear that Crooke's radiant matter, atoms, x-rays, radio waves, radioactivity, and the ether were any more real than ghosts, table tapping, and communicating with the dead; they all called into question established physics.

Crooke felt that scientists had a calling to investigate all these mysteries, and should avoid preconceptions about what was true or false. Others accused him of betraying his scientific calling and of being duped by spiritualists. Perhaps remembering that even the word "scientist" was unknown prior to the 1830's, they knew, nevertheless, that, if the history of the nine-

teenth century taught them anything, it was that there were also serious problems on the horizon of which they were completely unaware.

## 12.13 If You Can Spray Them, They Exist

Up through the seventeenth century, philosophy, especially the works of Aristotle, had colored the way scientists looked at the physical world. By the end of the nineteenth century, most scientists would have agreed that philosophy had been banished from science, that *metaphysics*, that is, statements that could not be empirically verified, had no place in science. But philosophy began to sneak back in, as questions about causality and the existence of objects we cannot see, such as atoms, started to be asked [5]. Most scientists are probably *realists*, believing that the objects they study have an existence independent of the instruments used to probe them. On the other side of the debate, *positivists*, or, at least, the more extreme positivists, hold that we have no way of observing an observer-independent reality, and therefore cannot verify that there is such a reality. Positivists hold that scientific theories are simply instruments used to hold together observed facts and make predictions. They do accept that the theories describe an *empirical* reality that is the same for all observers, but not a reality independent of observation. At first, scientists felt that it was safe for them to carry on without worrying too much about these philosophical points, but quantum theory would change things [98].

The idea that matter is composed of very small indivisible *atoms* goes back to the ancient Greeks. But it wasn't until after Einstein's 1905 paper on Brownian motion and subsequent experimental confirmations of his predictions that the actual existence of atoms was more or less universally accepted.

I recall reading somewhere about a conversation between a philosopher of science and an experimental physicist, in which the physicist was explaining how he sprayed an object with positrons. The philosopher then asked him if he really believed that positrons exist. The physicist answered, "If you can spray them, they exist."

## 12.14 What's Going On Here?

Experiments with cathode rays revealed that they were deflected by magnets, unlike any form of radiation similar to light, and unresponsive to gravity. Maybe they were very small electrically charged particles. In 1897 J.J. Thomson established that the cathode rays were, indeed, electrically charged particles, which he called *electrons*. For this discovery he was awarded the Nobel Prize in Physics in 1906. Perhaps there were two fundamental objects in nature, the atoms of materials and the electrons.

However, Volta's experiments suggested the electrons were within the material and involved in chemical reactions. In 1899 Thomson investigated the photo-electric effect and found that cathode rays could be produced by shining light on certain metals; the photo-electric effect revealed that electrons were inside the materials. Were they between the atoms, or inside the atoms? If they were within the atoms, perhaps their number and configuration could help explain Mendeleev's periodic table and the variety of elements found in nature.

In 1912, Max von Laue demonstrated that Röntgen's x-ray beams can be diffracted; this provided a powerful tool for determining the structure of crystals and molecules and later played an important role in the discovery of the double-helix structure of DNA. In 1923, the French physicist Louis de Broglie suggested that moving particles, such as electrons, should exhibit wave-like properties characterized by a wave-length. In particular, he suggested that beams of electrons sent through a narrow aperture could be diffracted. In 1937 G.P. Thomson, the son of J.J. Thomson, shared the Nobel Prize in Physics with Clinton Davisson for their work demonstrating that beams of electrons can be diffracted. As someone once put it, "The father won the prize for showing that electrons are particles, and the son won it for showing that they aren't." Some suggested that, since beams of electrons exhibited wave-like properties, they should give rise to the sort of interference effects Young had shown were exhibited by beams of light. The first laboratory experiment showing double-slit interference effects of beams of electrons was performed in 1989.

J.J. Thomson also discovered that the kinetic energy of the emitted electrons depended not at all on the intensity of the light, but only on its frequency. This puzzling aspect of the photo-electric effect prompted Einstein to consider the possibility that light is quantized, that is, it comes in small "packages", or *light quanta*, later called *photons*. Max Planck had earlier suggested that energy might be quantized, in order to explain the absence of the *ultraviolet catastrophe* in black-body radiation. It was his 1905 work on the photo-electric effect, not his work on special and general relativity, that eventually won for Einstein the 1921 Nobel Prize in Physics.

Einstein's 1905 paper that deals with the photo-electric effect is really a paper about the particle nature of light. But this idea met with great resistance, and it was made clear to Einstein that his prize was not for the whole paper, but for that part dealing with the photo-electric effect. He was even asked not to mention the particle nature of light in his Nobel speech.

Were the electrons the only sub-atomic particles? No, as Rutherford's discovery of the atomic nucleus in 1911 would reveal. And what is radioactivity, anyway? The new century was dawning, and all these questions were in the air. It was about 1900, Planck had just discovered the quantum theory, Einstein was in the patent office, where he would remain until

1909, Bohr and Schrödinger schoolboys, Heisenberg not yet born. A new scientific revolution was about to occur, and, as in 1800, nobody could have guessed what was coming next [118].

## 12.15 The Year of the Golden Eggs

As Rigden relates in [135], toward the end of his life Einstein looked back to 1905, when he was twenty-six, and told Leo Szilard, "They were the happiest years of my life. Nobody expected me to lay golden eggs." It is appropriate to end our story in 1905 because it was both an end and a beginning. In five great papers published in that year, Einstein solved several of the major outstanding problems that had worried physicists for years, but the way he answered them was revolutionary and began a whole new era of physics. After 1905 the development of electromagnetism merges with that of quantum mechanics, and becomes too big a story to relate here.

The problems that attracted Einstein involved apparent contradictions, and his answers were surprising. Is matter continuous or discrete? It is discrete; atoms do exist. Is light wave-like or particle-like? It is both. Are the laws of thermodynamics absolute or statistical? They are statistical. Are the laws of physics the same for observers moving with uniform velocity relative to one another? Yes; in particular, each will measure the speed of light to be the same. And, by the way, our notion of three-dimensional space and a separate dimension of time is wrong (special relativity), and gravity and acceleration are really the same thing (general relativity). Is inertial mass the same as gravitational mass? Yes. And what is mass, anyway? It is really energy, as $E = mc^2$ tells us.

## 12.16 Do Individuals Matter?

Our brief history of electromagnetism has focused on a handful of extraordinary people. But how important are individuals in the development of science, or in the course of history generally? An ongoing debate among those who study history is over the role of the Great Man [75]. On one side of the debate is the British writer and hero-worshipper Carlyle: "Universal history, the history of what man has accomplished in this world, is at bottom the History of the Great Men who have worked here." On the other side is the German political leader Bismarck: "The statesman's task is to hear God's footsteps marching through history, and to try to catch on to His coattails as He marches past."

If Mozart had never lived, nobody else would have composed his music. If Picasso had never lived, nobody else would have painted his pictures. If Winston Churchill had never lived, or had he died of his injuries when, in 1930, he was hit by a car on Fifth Avenue in New York City, western

Europe would probably be different today. If Hitler had died in 1930, when the car he was riding in was hit by a truck, recent history would certainly be different, in ways hard for us to imagine. But, I think the jury is still out on this debate, at least as it applies to science.

If Darwin had never lived, someone else would have published roughly the same ideas, at about the same time; in fact, Alfred Russel Wallace did just that. If Einstein had not lived, somebody else, maybe Poincaré, would have hit on roughly the same ideas, perhaps a bit later. Relativity would have been discovered by someone else. The fact that light behaves both like a wave and like a particle would have become apparent to someone else. The fact that atoms do really exist would have been demonstrated by someone else, although perhaps in a different way.

Nevertheless, just as Mozart's work is unique, even though it was obviously influenced by the times in which he composed and is clearly in the style of the late 18th century, Darwin's view of what he was doing differed somewhat from the view taken by Wallace, and Einstein's work reflected his own fascination with apparent contradiction and a remarkable ability, "to think outside the box", as the currently popular expression has it. Each of the people we have encountered in this brief history made a unique contribution, even though, had they not lived, others would probably have made their discoveries, one way or another.

People matter in another way, as well. Science is the work of individual people just as art, music and politics are. The book of nature, as some call it, is not easily read. Science is a human activity. Scientists are often mistaken and blind to what their training and culture prevent them from seeing. The history of the development of science is, like all history, our own story.

## 12.17 What's Next?

The twentieth century has taught us that all natural phenomena are based on two physical principles, quantum mechanics and relativity. The combination of special relativity and quantum mechanics led to a unification of three of the four fundamental forces of nature, electromagnetic force and the weak and strong nuclear forces, originally thought to be unrelated. The remaining quest is to combine quantum mechanics with general relativity, which describes gravity. Such a unification seems necessary if one is to solve the mysteries posed by *dark matter* and *dark energy* [16], which make up over three-quarters of the *stuff* of the universe, but of which nothing is known and whose existence can only be inferred from their gravitational effects. Perhaps what will be needed is a *paradigm shift*, to use Kuhn's popular phrase; perhaps the notion of a *fundamental particle*, or even of an *observer* will need to be abandoned.

The June 2010 issue of *Scientific American* contains an article called "Twelve events that will change everything". The article identifies twelve events, both natural and man-made, that could happen at any time and would transform society. It also rates the events in terms of how likely they are to occur: fusion energy (very unlikely); extraterrestrial intelligence, nuclear exchange, and asteroid collision (unlikely); deadly pandemic, room-temperature superconductors, and extra dimensions (50-50); cloning of a human, machine self-awareness, and polar meltdown (likely); and creation of life, and Pacific earthquake (almost certain). Our brief study of the history of electromagnetism should convince us that the event that will *really* change everything is not on this list nor on anyone else's list. As Brian Greene suggests [96], people in the year 2100 may look back on today as the time when the first primitive notions of parallel universes began to take shape.

## 12.18 Epilogue

As Butterfield points out in [22], science became modern in the period 1300 to 1800 not when experiment and observation replaced adherence to the authority of ancient philosophers, but when the experimentation was performed under the control of mathematics. New mathematical tools, logarithms, algebra, analytic geometry, and calculus, certainly played an important role, but so did mathematical thinking, measuring quantities, rather than speculating about qualities, idealizing and abstracting from a physical situation, and the like. Astronomy and mechanics were the first to benefit from this new approach. Paradoxically, our understanding of electromagnetism rests largely on a century or more of intuition, conjecture, experimentation and invention that was almost completely free of mathematics. To a degree, this was because the objects of interest, magnets and electricity, were close at hand and, increasingly, available for study. In contrast, Newton's synthesis of terrestrial and celestial gravitation was necessarily largely a mathematical achievement; observational data was available, but experimentation was not possible.

With Maxwell and the mathematicians, electromagnetism became a modern science. Now electromagnetism could be studied with a pencil and paper, as well as with generators. Consequences of the equations could be tested in the laboratory and used to advance technology. The incompleteness of the theory, with regard to the ether, the arrow of time, the finite speed of light, also served to motivate further theoretical and experimental investigation.

As electromagnetism, in particular, and physics, generally, became more mathematical, studies of the very small (nuclear physics), the very large (the universe), and the very long ago (cosmology) became possible. The

search for unifying theories of everything became mathematical studies, the consequences of the theories largely beyond observation [147].

In 2000 the mathematical physicist Ed Witten wrote a paper describing the physics of the century just ending [151]. Even the title is revealing; the quest is for *mathematical* understanding. He points out that, as physics became more mathematical in the first half of the twentieth century, with relativity and non-relativistic quantum mechanics, it had a broad influence on mathematics itself. The equations involved were familiar to the mathematicians of the day, even if the applications were not, and their use in physics prompted further mathematical development, and the emergence of new fields, such as functional analysis. In contrast, the physics of the second half of the century involves mathematics, principally quantum concepts applied to fields, not just particles, the foundations of which are not well understood by mathematicians. This is mathematics with which even the mathematicians are not familiar. Providing a mathematical foundation for the standard model for particle physics should keep the mathematicians of the next century busy for a while. The most interesting sentence in [151] is *The quest to understand string theory may well prove to be a central theme in physics of the twenty-first century.* Are physicists now just trying to understand their own mathematics, instead of the physical world?

# Chapter 13

# The Trans-Atlantic Cable (Chapters 4,12)

## 13.1 Introduction

In 1815, at the end of the war with England, the US was a developing country, with most people living on small farms, eating whatever they could grow themselves. Only those living near navigable water could market their crops. Poor transportation and communication kept them isolated. By 1848, at the end of the next war, this time with Mexico, things were different. The US was a transcontinental power, integrated by railroads, telegraph, steamboats, the Erie Canal, and innovations in mass production and agriculture. In 1828, the newly elected President, Andrew Jackson, arrived in Washington by horse-drawn carriage; he left in 1837 by train. The most revolutionary change was in communication, where the recent advances in understanding electromagnetism produced the telegraph. It wasn't long before efforts began to lay a telegraph cable under the Atlantic Ocean, even though some wondered what England and the US could possibly have to say to one another.

The laying of the trans-Atlantic cable was, in many ways, the 19th century equivalent of landing a man on the moon, involving, as it did, considerable expense, too frequent failure, and a level of precision in engineering design and manufacturing never before attempted. From a scientific perspective, it was probably more difficult, given that the study of electromagnetism was in its infancy at the time.

Early on, Faraday and others worried that sending a message across a vast distance would take a long time, but they reasoned, incorrectly, that this would be similar to filling a very long hose with water. What they did not realize initially was that, as William Thomson was to discover,

the transmission of a pulse through an undersea cable was described more by a heat equation than a wave equation. This meant that a signal that started out as a sharp pulse would be spread out as time went on, making communication extremely slow. The problem was the increased capacitance with the ground.

Somewhat later, Oliver Heaviside realized that, when all four of the basic elements of the electrical circuit, the inductance, the resistance, the conductance to the ground and the capacitance to the ground, were considered together, it might be possible to adjust these parameters, in particular, to increase the inductance, so as to produce undistorted signals. Heaviside died in poverty, but his ideas eventually were adopted.

In 1859 Queen Victoria sent President Buchanan a 99 word greeting using an early version of the cable, but the message took over sixteen hours to be received. By 1866 one could transmit eight words a minute along a cable that stretched from Ireland to Newfoundland, at a cost of about 1500 dollars per word in today's money. With improvements in insulation, using gutta percha, a gum from a tropical tree also used to make golf balls, and the development of magnetic alloys that increased the inductance of the cable, messages could be sent faster and more cheaply.

In this chapter we survey the development of the mathematics of the problem. We focus, in particular, on the partial differential equations that were used to describe the transmission problem. What we give here is a brief glimpse; more detailed discussion of this problem is found in the books by Körner [109], Gonzalez-Velasco [91], and Wylie [154].

## 13.2 The Electrical Circuit ODE

We begin with the ordinary differential equation that describes the horizontal motion of a block of wood attached to a spring. We let $x(t)$ be the position of the block relative to the equilibrium position $x = 0$, with $x(0)$ and $x'(0)$ denoting the initial position and velocity of the block. When an external force $f(t)$ is imposed, a portion of this force is devoted to overcoming the inertia of the block, a portion to compressing or stretching the spring, and the remaining portion to resisting friction. Therefore, the differential equation describing the motion is

$$mx''(t) + ax'(t) + kx(t) = f(t), \tag{13.1}$$

where $m$ is the mass of the block, $a$ the coefficient of friction, and $k$ the spring constant.

The charge $Q(t)$ deposited on a capacitor in an electrical circuit due to an imposed electromotive force $E(t)$ is similarly described by the ordinary

differential equation

$$LQ''(t) + RQ'(t) + \frac{1}{C}Q(t) = E(t). \tag{13.2}$$

The first term, containing the inductance coefficient $L$, describes the portion of the force $E(t)$ devoted to overcoming the effect of a change in the current $I(t) = Q'(t)$; here $L$ is analogous to the mass $m$. The second term, containing the resistance coefficient $R$, describes that portion of the force $E(t)$ needed to overcome resistance to the current $I(t)$; now $R$ is analogous to the friction coefficient $a$. Finally, the third term, containing the reciprocal of the capacitance $C$, describes the portion of $E(t)$ used to store charge on the capacitor; now $\frac{1}{C}$ is analogous to $k$, the spring constant.

## 13.3 The Telegraph Equation

The objective here is to describe the behavior of $u(x,t)$, the voltage at location $x$ along the cable, at time $t$. In the beginning, it was believed that the partial differential equation describing the voltage would be the wave equation

$$u_{xx} = \alpha^2 u_{tt}.$$

If this were the case, an initial pulse

$$E(t) = H(t) - H(t-T)$$

would move along the cable undistorted; here $H(t)$ is the Heaviside function that is zero for $t < 0$ and one for $t \geq 0$. Thomson (later Sir William Thomson, and even later, Lord Kelvin) thought otherwise.

Thomson argued that there would be a voltage drop over an interval $[x, x+\Delta x]$ due to resistance to the current $i(x,t)$ passing through the cable, so that

$$u(x+\Delta x, t) - u(x,t) = -Ri(x,t)\Delta x,$$

and so

$$\frac{\partial u}{\partial x} = -Ri.$$

He also argued that there would be capacitance to the ground, made more significant under water. Since the apparent change in current due to the changing voltage across the capacitor is

$$i(x+\Delta x, t) - i(x,t) = -Cu_t(x,t)\Delta x,$$

we have

$$\frac{\partial i}{\partial x} = -C\frac{\partial u}{\partial t}.$$

Eliminating the $i(x,t)$, we can write

$$u_{xx}(x,t) = CRu_t(x,t), \tag{13.3}$$

which is the heat equation, not the wave equation.

## 13.4 Consequences of Thomson's Model

To see what Thomson's model predicts, we consider the following problem. Suppose we have a semi-infinite cable, that the voltage is $u(x,t)$ for $x \geq 0$, and $t \geq 0$, and that $u(0,t) = E(t)$. Let $U(x,s)$ be the Laplace transform of $u(x,t)$, viewed as a function of $t$. Then, from Thomson's model we have

$$U(x,s) = \mathcal{L}(E)(s)e^{-\sqrt{CRs}x},$$

where $\mathcal{L}(E)(s)$ denotes the Laplace transform of $E(t)$. Since $U(x,s)$ is the product of two functions of $s$, the convolution theorem applies. But first, it is helpful to find out which function has for its Laplace transform the function $e^{-\alpha x\sqrt{s}}$. The answer comes from the following fact: the function

$$be^{-b^2/4t}/2\sqrt{\pi}t^{3/2}$$

has for its Laplace transform the function $e^{-b\sqrt{s}}$. Therefore, we can write

$$u(x,t) = \frac{\sqrt{CR}x}{2\sqrt{\pi}} \int_0^t E(t-\tau)\frac{e^{-CRx^2/4\tau}}{\tau\sqrt{\tau}}d\tau.$$

Now we consider two special cases.

### 13.4.1 Special Case 1: $E(t) = H(t)$

Suppose now that $E(t) = H(t)$, the Heaviside function. Using the substitution

$$z = CRx^2/4\tau,$$

we find that

$$u(x,t) = 1 - \frac{2}{\sqrt{\pi}} \int_0^{\sqrt{CR}x/2\sqrt{\pi}} e^{-z^2}dz. \tag{13.4}$$

The function

$$\text{erf}(r) = \frac{2}{\sqrt{\pi}} \int_0^r e^{-z^2}dz$$

is the well known *error function*, so we can write

$$u(x,t) = 1 - \text{erf}\Big(\frac{\sqrt{CR}x}{2\sqrt{t}}\Big). \tag{13.5}$$

### 13.4.2 Special Case 2: $E(t) = H(t) - H(t-T)$

Now suppose that $E(t)$ is the pulse $H(t) - H(t-T)$. Using the results from the previous subsection, we find that, for $t > T$,

$$u(x,t) = \operatorname{erf}\Big(\frac{\sqrt{CR}x}{2\sqrt{t-T}}\Big) - \operatorname{erf}\Big(\frac{\sqrt{CR}x}{2\sqrt{t}}\Big). \tag{13.6}$$

For fixed $x$, $u(x,t)$ is proportional to the area under the function $e^{-z^2}$, over an interval that, as time goes on, moves steadily to the left and decreases in length. For small $t$ the interval involves only large $z$, where the function $e^{-z^2}$ is nearly zero and the integral is nearly zero. As $t$ increases, the interval of integration moves to the left, so that the integrand grows larger, but the length of the interval grows smaller. The net effect is that the voltage at $x$ increases gradually over time, and then decreases gradually; the sharp initial pulse is smoothed out in time.

## 13.5 Heaviside to the Rescue

It seemed that Thomson had solved the mathematical problem and discovered why the behavior was not wave-like. Since it is not really possible to reduce the resistance along the cable, and capacitance to the ground would probably remain a serious issue, particularly under water, it appeared that little could be done to improve the situation. But Heaviside had a solution.

Heaviside argued that Thomson had ignored two other circuit components, the leakage of current to the ground, and the self-inductance of the cable. He revised Thomson's equations, obtaining

$$u_x = -Li_t - Ri,$$

and

$$i_x = -Cu_t - Gu,$$

where $L$ is the inductance and $G$ is the coefficient of leakage of current to the ground. The partial differential equation governing $u(x,t)$ now becomes

$$u_{xx} = LCu_{tt} + (LG + RC)u_t + RGu, \tag{13.7}$$

which is the formulation used by Kirchhoff. As Körner remarks, never before had so much money been riding on the solution of one partial differential equation.

### 13.5.1 A Special Case: $G = 0$

If we take $G = 0$, thereby assuming that no current passes into the ground, the partial differential equation becomes

$$u_{xx} = LCu_{tt} + RCu_t, \tag{13.8}$$

or

$$\frac{1}{CL}u_{xx} = u_{tt} + \frac{R}{L}u_t. \tag{13.9}$$

If $R/L$ could be made small, we would have a wave equation again, but with a propagation speed of $1/\sqrt{CL}$. This suggested to Heaviside that one way to obtain undistorted signaling would be to increase $L$, since we cannot realistically hope to change $R$. He argued for years for the use of cables with higher inductance, which eventually became the practice, helped along by the invention of new materials, such as magnetic alloys, that could be incorporated into the cables.

### 13.5.2 Another Special Case

Assume now that $E(t)$ is the pulse. Applying the Laplace transform method described earlier to Equation (13.7), we obtain

$$U_{xx}(x,s) = (Cs+G)(Ls+R)U(x,s) = \lambda^2 U(x,s),$$

from which we get

$$U(x,s) = A(s)e^{\lambda x} + \Big(\frac{1}{s}(1-e^{-Ts}) - A(s)\Big)e^{-\lambda x}.$$

If it happens that $GL = CR$, we can solve easily for $\lambda$:

$$\lambda = \sqrt{CL}s + \sqrt{GR}.$$

Then we have

$$U(x,s) = e^{-\sqrt{GR}x}\frac{1}{s}(1-e^{-Ts})e^{-\sqrt{CL}xs},$$

so that

$$u(x,t) = e^{-\sqrt{GR}x}\Big(H(t-x\sqrt{CL}) - H(t-T-x\sqrt{CL})\Big). \tag{13.10}$$

This tells us that we have an undistorted pulse that arrives at the point $x$ at the time $t = x\sqrt{CL}$.

In order to have $GL = CR$, we need $L = CR/G$. Since $C$ and $R$ are more or less fixed, and $G$ is typically reduced by insulation, $L$ will need to be large. Again, this argues for increasing the inductance in the cable.

# Chapter 14

# Hermite's Equations and Quantum Mechanics (Chapter 10,11)

## 14.1 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

$$i\hbar\frac{\partial\psi(x,t)}{\partial t} = -\frac{\hbar}{2m}\frac{\partial^2\psi(x,t)}{\partial x^2} + V(x,t)\psi(x,t), \tag{14.1}$$

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.

## 14.2 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 (14.1) by separating the variables; we take $\psi(x,t) = f(t)g(x)$ and insert this product into Equation (14.1).

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*

$$-\frac{\hbar}{2m}g''(x) + V(x)g(x) = Eg(x). \tag{14.2}$$

An important special case is the harmonic oscillator.

## 14.3 The Harmonic Oscillator

The case of the *harmonic oscillator* corresponds to the potential $V(x) = \frac{1}{2}kx^2$.

### 14.3.1 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.

### 14.3.2 Back to the Harmonic Oscillator

When the potential function is $V(x) = \frac{1}{2}kx^2$, Equation (14.2) becomes

$$\frac{\hbar}{2m}g''(x) + (E - \frac{1}{2}kx^2)g(x) = 0, \tag{14.3}$$

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

$$\frac{d^2g}{du^2} + (\epsilon - u^2)g = 0. \tag{14.4}$$

Equation (14.4) 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 (14.3) 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.

## 14.4 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 (14.1), 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:

$$i\hbar\frac{\partial\psi}{\partial t} = \frac{\hbar c}{i}\Big(\alpha_1\frac{\partial\psi}{\partial x_1} + \alpha_2\frac{\partial\psi}{\partial x_2} + \alpha_3\frac{\partial\psi}{\partial x_3}\Big) + \alpha_4 mc^2\psi. \tag{14.5}$$

Here the $\alpha_i$ 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 a 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.

# Bibliography


[1] 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.

[2] Anderson, A. and Kak, A. (1984) "Simultaneous algebraic reconstruction technique (SART): a superior implementation of the ART algorithm." *Ultrasonic Imaging* **6**, pp. 81–94.

[3] Aubin, J.-P., (1993) *Optima and Equilibria: An Introduction to Nonlinear Analysis*, Springer-Verlag.

[4] Auslander, A., and Teboulle, M. (2006) "Interior gradient and proximal methods for convex and conic optimization." *SIAM Journal on Optimization*, **16(3)**, pp. 697–725.

[5] Baggott, J. (1992) *The Meaning of Quantum Theory*, Oxford University Press.

[6] Barrett, H., White, T., and Parra, L. (1997) "List-mode likelihood." *J. Opt. Soc. Am. A* **14**, pp. 2914–2923.

[7] Bauschke, H., and Borwein, J. (1993) "On the convergence of von Neumann's alternating projection algorithm for two sets." *Set-Valued Analysis* **1**, pp. 185–212.

[8] Bauschke, H., and Borwein, J. (1996) "On projection algorithms for solving convex feasibility problems." *SIAM Review*, **38 (3)**, pp. 367–426.

[9] Bauschke, H., and Borwein, J. (2001) "Joint and separate convexity of the Bregman distance." in *Inherently Parallel Algorithms in Feasibility and Optimization and their Applications*, edited by D. Butnariu, Y. Censor and S. Reich, pp. 23–36, Studies in Computational Mathematics 8. Amsterdam: Elsevier Publ.

[10] Bauschke, H., and Combettes, P. (2003) "Iterating Bregman retractions." *SIAM Journal on Optimization*, **13**, pp. 1159–1173.

[11] Bauschke, H., Combettes, P., and Noll, D. (2006) "Joint minimization with alternating Bregman proximity operators." *Pacific Journal of Optimization*, **2**, pp. 401–424.

[12] Bertero, M., and Boccacci, P. (1998) *Introduction to Inverse Problems in Imaging* Bristol, UK: Institute of Physics Publishing.

[13] Bochner, S. (1966) *The Role of Mathematics in the Rise of Science*. Princeton University Press.

[14] Boyles, R. (1983) "On the convergence of the EM algorithm." *Journal of the Royal Statistical Society B*, **45**, pp. 47–50.

[15] 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.

[16] Brockman, M. (2009) *What's Next? Dispatches on the Future of Science*, Vintage Books, New York.

[17] 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.

[18] Burg, J. (1967) "Maximum entropy spectral analysis." *paper presented at the 37th Annual SEG meeting, Oklahoma City, OK.*

[19] Burg, J. (1972) "The relationship between maximum entropy spectra and maximum likelihood spectra." *Geophysics* **37**, pp. 375–376.

[20] Burg, J. (1975) *Maximum Entropy Spectral Analysis*, Ph.D. dissertation, Stanford University.

[21] Butnariu, D., Byrne, C., and Censor, Y. (2003) "Redundant axioms in the definition of Bregman functions." *Journal of Convex Analysis*, **10**, pp. 245–254.

[22] Butterfield, H. (1957) *The Origins of Modern Science: 1300–1800*, Free Press Paperback (MacMillan Co.).

[23] Byrne, C. and Fitzgerald, R. (1979) "A unifying model for spectrum estimation." In *Proceedings of the RADC Workshop on Spectrum Estimation*, Griffiss AFB, Rome, NY, October.

[24] Byrne, C. and Fitzgerald, R. (1982) "Reconstruction from partial information, with applications to tomography." *SIAM J. Applied Math.* **42(4)**, pp. 933–940.

[25] 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.

[26] 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.

[27] 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.

[28] Byrne, C. and Fiddy, M. (1987) "Estimation of continuous object distributions from Fourier magnitude measurements." *JOSA A* **4**, pp. 412–417.

[29] Byrne, C. and Fiddy, M. (1988) "Images as power spectra; reconstruction as Wiener filter approximation." *Inverse Problems* **4**, pp. 399–409.

[30] Byrne, C. (1993) "Iterative image reconstruction algorithms based on cross-entropy minimization." *IEEE Transactions on Image Processing* **IP-2**, pp. 96–103.

[31] 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.

[32] 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.

[33] Byrne, C. (1996) "Block-iterative methods for image reconstruction from projections." *IEEE Transactions on Image Processing* **IP-5**, pp. 792–794.

[34] Byrne, C. (1997) "Convergent block-iterative algorithms for image reconstruction from inconsistent data." *IEEE Transactions on Image Processing* **IP-6**, pp. 1296–1304.

[35] 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.

[36] Byrne, C. (1998) "Iterative algorithms for deblurring and deconvolution with constraints." *Inverse Problems*, **14**, pp. 1455–1467.

[37] Byrne, C. (2000) "Block-iterative interior point optimization methods for image reconstruction from limited data." *Inverse Problems* **16**, pp. 1405–1419.

[38] Byrne, C. (2001) "Bregman-Legendre multi-distance projection algorithms for convex feasibility and optimization." in *Inherently Parallel Algorithms in Feasibility and Optimization and their Applications*, edited by D. Butnariu, Y. Censor and S. Reich, pp. 87-100, Studies in Computational Mathematics 8. Amsterdam: Elsevier Publ.

[39] Byrne, C. (2001) "Likelihood maximization for list-mode emission tomographic image reconstruction." *IEEE Transactions on Medical Imaging* **20(10)**, pp. 1084–1092.

[40] 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.

[41] Byrne, C. (2002) "Iterative oblique projection onto convex sets and the split feasibility problem." *Inverse Problems* **18**, pp. 441–453.

[42] Byrne, C. (2004) "A unified treatment of some iterative algorithms in signal processing and image reconstruction." *Inverse Problems* **20**, pp. 103–120.

[43] Byrne, C. (2005) "Choosing parameters in block-iterative or ordered-subset reconstruction algorithms." *IEEE Transactions on Image Processing*, **14 (3)**, pp. 321–327.

[44] Byrne, C. (2005) *Signal Processing: A Mathematical Approach*, AK Peters, Publ., Wellesley, MA.

[45] Byrne, C. (2007) *Applied Iterative Methods*, AK Peters, Publ., Wellesley, MA.

[46] Byrne, C. (2008) "Sequential unconstrained minimization algorithms for constrained optimization." *Inverse Problems*, **24(1)**, article no. 015013.

[47] Byrne, C. (2009) "Block-iterative algorithms." *International Transactions in Operations Research*, **16(4)**, pp. 427–463.

[48] 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)**, pp. 465–479.

[49] Byrne, C. (2009) *Applied and Computational Linear Algebra: A First Course*, available as a pdf file at my web site.

[50] Byrne, C. (2011) *A First Course in Optimization*, available as a pdf file at my web site.

[51] Byrne, C. "Alternating Minimization and Alternating Projection." *preprint*.

[52] Byrne, C. (2012) "Alternating and sequential unconstrained minimization algorithms." submitted.

[53] Byrne, C., and Eggermont, P. (2011) "EM Algorithms." in *Handbook of Mathematical Methods in Imaging*, Otmar Scherzer, ed., Springer-Science.

[54] Byrne, C., and Eggermont, P. (2012) "EM algorithms." in preparation.

[55] 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 .

[56] Censor, Y. and Elfving, T. (1994) "A multi-projection algorithm using Bregman projections in a product space." *Numerical Algorithms*, **8**, pp. 221–239.

[57] 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), 2353-2365.

[58] Censor, Y., Elfving, T., Kopf, N., and Bortfeld, T. (2005) "The multiple-sets split feasibility problem and its application for inverse problems." *Inverse Problems*, **21** , pp. 2071-2084.

[59] Censor, Y. and Segman, J. (1987) "On block-iterative maximization." *J. of Information and Optimization Sciences* **8**, pp. 275–291.

[60] Censor, Y., and Zenios, S.A. (1992) "Proximal minimization algorithm with $D$-functions." *Journal of Optimization Theory and Applications*, **73(3)**, pp. 451–464.

[61] Censor, Y. and Zenios, S.A. (1997) *Parallel Optimization: Theory, Algorithms and Applications*. New York: Oxford University Press.

[62] 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.

[63] 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.

[64] Censor, Y., Eggermont, P.P.B., and Gordon, D. (1983) "Strong under-relaxation in Kaczmarz's method for inconsistent systems." *Numerische Mathematik* **41**, pp. 83–92.

[65] Cheney, W., and Goldstein, A. (1959) "Proximity maps for convex sets." *Proc. Amer. Math. Soc.*, **10**, pp. 448–450.

[66] Cimmino, G. (1938) "Calcolo approssimato per soluzioni dei sistemi di equazioni lineari." *La Ricerca Scientifica XVI, Series II, Anno IX* **1**, pp. 326–333.

[67] Combettes, P., and Wajs, V. (2005) "Signal recovery by proximal forward-backward splitting." *Multiscale Modeling and Simulation*, **4(4)**, pp. 1168–1200.

[68] Csiszár, I. (1975) "I-divergence geometry of probability distributions and minimization problems." *The Annals of Probability* **3(1)**, pp. 146–158.

[69] Csiszár, I. (1989) "A geometric interpretation of Darroch and Ratcliff's generalized iterative scaling." *The Annals of Statistics* **17(3)**, pp. 1409–1413.

[70] Csiszár, I. and Tusnády, G. (1984) "Information geometry and alternating minimization procedures." *Statistics and Decisions* **Supp. 1**, pp. 205–237.

[71] Darroch, J. and Ratcliff, D. (1972) "Generalized iterative scaling for log-linear models." *Annals of Mathematical Statistics* **43**, pp. 1470–1480.

[72] 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.

[73] 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)**.

[74] 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.

[75] Diamond, J. (1997) *Guns, Germs, and Steel*, Norton, Publ.

[76] Dolidze, Z.O. (1982) "Solution of variational inequalities associated with a class of monotone maps." *Ekonomika i Matem. Metody* **18 (5)**, pp. 925–927 (in Russian).

[77] Donoho, D. (2006) "Compressed sampling" *IEEE Transactions on Information Theory*, **52 (4)**. (download preprints at http://www.stat.stanford.edu/ donoho/Reports).

[78] Eggermont, P.P.B., LaRiccia, V.N. (1995) "Smoothed maximum likelihood density estimation for inverse problems." Annals of Statistics **23**, pp. 199–220.

[79] Eggermont, P., and LaRiccia, V. (2001) *Maximum Penalized Likelihood Estimation, Volume I: Density Estimation*. New York: Springer.

[80] Erdogan, H., and Fessler, J. (1999) "Fast monotonic algorithms for transmission tomography" *IEEE Transactions on Medical Imaging*, **18(9)**, pp. 801–814.

[81] Everitt, B. and Hand, D. (1981) *Finite Mixture Distributions* London: Chapman and Hall.

[82] Fara, P. (2009) *Science: A Four Thousand Year History*, Oxford University Press.

[83] 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.

[84] Feynman, R., Leighton, R., and Sands, M. (1963) *The Feynman Lectures on Physics, Vol. 1*. Boston: Addison-Wesley.

[85] Fiacco, A., and McCormick, G. (1990) *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*. Philadelphia, PA: SIAM Classics in Mathematics (reissue).

[86] Fiddy, M. (2008) *private communication*.

[87] 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.

[88] Gerchberg, R. W. (1974) "Super-restoration through error energy reduction." *Optica Acta* **21**, pp. 709–720.

[89] Gill, P., Murray, W., Saunders, M., Tomlin, J., and Wright, M. (1986) "On projected Newton barrier methods for linear programming and an equivalence to Karmarkar's projective method." *Mathematical Programming*, **36**, pp. 183–209.

[90] Golshtein, E., and Tretyakov, N. (1996) *Modified Lagrangians and Monotone Maps in Optimization*. New York: John Wiley and Sons, Inc.

[91] Gonzalez-Velasco, E. (1996) *Fourier Analysis and Boundary Value Problems*. Academic Press.

[92] Gonzalez-Velasco, E. (2008) *personal communication*.

[93] 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.

[94] Graham-Eagle, J. (2008) unpublished notes in applied mathematics.

[95] Green, P. (1990) "Bayesian reconstructions from emission tomography data using a modified EM algorithm." *IEEE Transactions on Medical Imaging* **9**, pp. 84–93.

[96] Greene, B. (2011) *The Hidden Reality: Parallel Universes and the Deep Laws of the Cosmos*. New York: Vintage Books.

[97] 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.

[98] Heisenberg, W. (1958) *Physics and Philosophy*, Harper Torchbooks.

[99] Herman, G.T., Censor, Y., Gordon, D., and Lewitt, R. (1985) "Comment (on the paper [149])." *Journal of the American Statistical Association* **80**, pp. 22–25.

[100] Herman, G. T. (1999) *private communication*.

[101] Hogg, R., McKean, J., and Craig, A. (2004) *Introduction to Mathematical Statistics*, 6th edition, Prentice Hall.

[102] 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.

[103] Hudson, M., Hutton, B., and Larkin, R. (1992) "Accelerated EM reconstruction using ordered subsets." *Journal of Nuclear Medicine*, **33**, p.960.

[104] 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.

[105] 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.

[106] Kaczmarz, S. (1937) "Angenäherte Auflösung von Systemen linearer Gleichungen." *Bulletin de l'Academie Polonaise des Sciences et Lettres* **A35**, pp. 355–357.

[107] Karmarkar, N. (1984) "A new polynomial-time algorithm for linear programming." *Combinatorica*, **4**, pp. 373–395.

[108] Koestler, A. (1959) *The Sleepwalkers: A History of Man's Changing Vision of the Universe*, Penguin Books.

[109] Körner, T. (1988) *Fourier Analysis*. Cambridge, UK: Cambridge University Press.

[110] Krasnosel'skiĭ, M. (1955) "Two remarks on the method of successive approximations" (in Russian). *Uspekhi Mathematicheskikh Nauk*, **10**, pp. 123–127.

[111] Kullback, S. and Leibler, R. (1951) "On information and sufficiency." *Annals of Mathematical Statistics* **22**, pp. 79–86.

[112] Landweber, L. (1951) "An iterative formula for Fredholm integral equations of the first kind." *Amer. J. of Math.* **73**, pp. 615–624.

[113] Lange, K. and Carson, R. (1984) "EM reconstruction algorithms for emission and transmission tomography." *Journal of Computer Assisted Tomography* **8**, pp. 306–316.

[114] 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.

[115] 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.

[116] 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.

[117] Lindberg, D. (1992) *The Beginnings of Western Science*, University of Chicago Press.

[118] Lindley, D. (2007) *Uncertainty: Einstein, Heisenberg, Bohr, and the Struggle for the Soul of Science*, Doubleday.

[119] Mann, W. (1953) "Mean value methods in iteration." *Proceedings of the American Mathematical Society*, **4**, pp. 506–510.

[120] McLachlan, G.J. and Krishnan, T. (1997) *The EM Algorithm and Extensions*. New York: John Wiley and Sons, Inc.

[121] 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.

[122] Meng, X., and Pedlow, S. (1992) "EM: a bibliographic review with missing articles." *Proceedings of the Statistical Computing Section, American Statistical Association*, American Statistical Association, Alexandria, VA.

[123] Meng, X., and van Dyk, D. (1997) "The EM algorithm- An old folksong sung to a fast new tune." *J. R. Statist. Soc. B*, **59(3)**, pp. 511–567.

[124] Moreau, J.-J. (1962) "Fonctions convexes duales et points proximaux dans un espace hilbertien." *C.R. Acad. Sci. Paris Sér. A Math.*, **255**, pp. 2897–2899.

[125] Moreau, J.-J. (1963) "Propriétés des applications 'prox'." *C.R. Acad. Sci. Paris Sér. A Math.*, **256**, pp. 1069–1071.

[126] Moreau, J.-J. (1965) "Proximité et dualité dans un espace hilbertien." *Bull. Soc. Math. France*, **93**, pp. 273–299.

[127] Muller, R. (2008) *Physics for Future Presidents: the Science Behind the Headlines*, Norton.

[128] 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.

[129] Nash, S. and Sofer, A. (1996) *Linear and Nonlinear Programming*. New York: McGraw-Hill.

[130] Nesterov, Y., and Nemirovski, A. (1994) *Interior-Point Polynomial Algorithms in Convex Programming*. Philadelphia, PA: SIAM Studies in Applied Mathematics.

[131] Papoulis, A. (1975) "A new algorithm in spectral analysis and band-limited extrapolation." *IEEE Transactions on Circuits and Systems* **22**, pp. 735–742.

[132] 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.

[133] Redner, R., and Walker, H. (1984) "Mixture Densities, Maximum Likelihood and the EM Algorithm." *SIAM Review*, **26(2)**, pp. 195–239.

[134] Renegar, J. (2001) *A Mathematical View of Interior-Point Methods in Convex Optimization*. Philadelphia, PA: SIAM (MPS-SIAM Series on Optimization).

[135] Rigden, J. (2005) *Einstein 1905: The Standard of Greatness*. Harvard University Press.

[136] Rockafellar, R. (1970) *Convex Analysis*. Princeton, NJ: Princeton University Press.

[137] 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.

[138] Schmidlin, P. (1972) "Iterative separation of sections in tomographic scintigrams." *Nucl. Med.* **15(1)**.

[139] Shepp, L., and Vardi, Y. (1982) "Maximum likelihood reconstruction for emission tomography." *IEEE Transactions on Medical Imaging*, **MI-1**, pp. 113–122.

[140] 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.

[141] 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.

[142] Shieh, M., and Byrne, C. (2006) "Image reconstruction from limited Fourier data." *Journal of the Optical Society of America, A*, **23(11)**, pp. 2732–2736.

[143] Silverman, B., Jones, M., Wilson, J., and Nychka, D. (1990) "A smoothed EM approach to indirect estimation problems, with particular reference to stereology and emission tomography (with discussion)." *Journal of the Royal Statistical Society* **B 52**, pp. 271–324.

[144] Simmons, G. (1972) *Differential Equations, with Applications and Historical Notes*. New York: McGraw-Hill.

[145] Smith, C. Ray and Grandy, W.T., editors (1985) *Maximum-Entropy and Bayesian Methods in Inverse Problems*. Dordrecht: Reidel Publ.

[146] Smith, C. Ray and Erickson, G., editors (1987) *Maximum-Entropy and Bayesian Spectral Analysis and Estimation Problems*. Dordrecht: Reidel Publ.

[147] Smolin, L. (2006) *The Trouble with Physics*, Houghton Mifflin.

[148] Teboulle, M. (1992) "Entropic proximal mappings with applications to nonlinear programming." *Mathematics of Operations Research*, **17(3)**, pp. 670–690.

[149] 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.

[150] Wernick, M. and Aarsvold, J., editors (2004) *Emission Tomography: The Fundamentals of PET and SPECT*. San Diego: Elsevier Academic Press.

[151] Witten, E. (2002) "Physical law and the quest for mathematical understanding." *Bulletin of the American Mathematical Society*, **40(1)**, pp. 21–29.

[152] Wright, M. (2005) "The interior-point revolution in optimization: history, recent developments, and lasting consequences." *Bulletin (New Series) of the American Mathematical Society*, **42(1)**, pp. 39–56.

[153] Wu, C.F.J. (1983) "On the convergence properties of the EM algorithm." *Annals of Statistics*, **11**, pp. 95–103.

[154] Wylie, C.R. (1966) *Advanced Engineering Mathematics*. New York: McGraw-Hill.

[155] Yang, Q. (2004) "The relaxed CQ algorithm solving the split feasibility problem." *Inverse Problems*, **20**, pp. 1261–1266.

# Index