Charles L. Byrne
Department of Mathematical Sciences
University of Massachusetts Lowell
December 8, 2014

Lecture Notes on Iterative Optimization Algorithms

Contents

Preface

I wrote these notes in order to help me discover how best to present the circle of definitions and propositions that center on the Baillon–Haddad and Krasnosel'skii-Mann-Opial Theorems in iterative optimization. This short booklet is, for the most part, an abbreviated version of my book *Iterative Optimization in Inverse Problems*, published in January, 2014, by CRC Press. Some of the results contained here are new, particularly those pertaining to the Baillon–Haddad Theorem. My articles listed in the bibliography can be downloaded from my website, http://faculty.uml.edu/cbyrne/cbyrne.html.

Chapter 1

Overview and Examples

1.1 Overview

The basic problem we consider in these notes is to minimize a function $f : X \to \mathbb{R}$ over x in $C \subseteq X$, where X is an arbitrary nonempty set. Until it is absolutely necessary, we shall not impose any structure on X or on f. One reason for avoiding structure on X and f is that we can actually achieve something interesting without it. The second reason is that when we do introduce structure, it will not necessarily be that of a metric space; for instance, cross-entropy and other Bregman distances play an important role in some of the iterative optimization algorithms I discuss in these notes.

We investigate two classes of iterative optimization methods: sequential

auxiliary-function (AF) methods; and fixed-point (FP) methods. As we shall see, there is some overlap between these two classes of methods. As is appropriate for an overview, in this chapter we make a number of assertions without providing proofs. Proofs of most of these assertions will be given in subsequent chapters.

1.2 Auxiliary-Function Methods

For $k = 1, 2, ...$ we minimize the function

$$G_k(x) = f(x) + g_k(x) \tag{1.1}$$

over x in X to get x^k. We shall say that the functions $g_k(x)$ are *auxiliary functions* if they have the properties $g_k(x) \geq 0$ for all $x \in X$, and $g_k(x^{k-1}) = 0$. We then say that the sequence $\{x^k\}$ has been generated by an *auxiliary-function* (AF) method. We then have the following result.

Proposition 1.1 *If the sequence $\{x^k\}$ is generated by an AF method, then the sequence $\{f(x^k)\}$ is nonincreasing.*

Proof: We have

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

$$\geq G_k(x^k) = f(x^k) + g_k(x^k) \geq f(x^k),$$

so $f(x^{k-1}) \geq f(x^k)$. ∎

In order to have the sequence $\{f(x^k\}$ converging to $\beta = \inf\{f(x)|x \in C\}$ we need to impose an additional property on the $g_k(x)$. We shall return to this issue later in this chapter.

Perhaps the best known examples of AF methods are the *sequential unconstrained minimization* (SUM) methods discussed by Fiacco and McCormick in their classic book [94]. They focus on barrier-function and penalty-function algorithms, in which the auxiliary functions are introduced to incorporate the constraint that f is to be minimized over C. In [94] barrier-function methods are called *interior-point methods*, while penalty-function methods are called *exterior-point methods*.

A barrier function has the value $+\infty$ for x not in C, while the penalty function is zero on C and positive off of C. In more general AF methods, we may or may not have $C = X$. If C is a proper subset of X, we can replace the function $f(x)$ with $f(x) + \iota_C(x)$, where $\iota_C(x)$ takes on the value zero for x in C and the value $+\infty$ for x not in C; then the $g_k(x)$ need not involve C.

Prior to the 1980's linear programming and nonlinear programming were viewed as separate subjects. In 1984 Karmarkar published his polynomial-time interior-point algorithm for linear programming [113]. This event created much excitement at the time, even making the front page of the New York Times. Although it appears now that some claims for this algorithm were overstated, the arrival of the algorithm served to revive interest in interior-point methods and to bring linear and nonlinear programming closer together.

As we shall see, in addition to incorporating the constraint set C, the $g_k(x)$ can be selected to make the computations simpler; sometimes we select the $g_k(x)$ so that x^k can be expressed in closed form. However, in the most general, non-topological case, we are not concerned with calculational issues involved in finding x^k. Our objective is to select the $g_k(x)$ so that the sequence $\{f(x^k)\}$ converges to $\beta = \inf\{f(x), x \in C\}$. We begin with two simple examples of the use of barrier functions.

1.2.1 Barrier-Function Methods: An Example

Our first problem is to minimize the function $f(x) = f(x_1, x_2) = x_1^2 + x_2^2$, subject to $x_1 + x_2 \geq 1$. Here $X = \mathbb{R}^2$, the function $f(x)$ is continuous, indeed, differentiable, and the set $C = \{x | x_1 + x_2 \geq 1\}$ is closed and convex. For each k we minimize the function

$$B_k(x) = f(x) + \frac{1}{k} b(x) = x_1^2 + x_2^2 - \frac{1}{k} \log(x_1 + x_2 - 1) \tag{1.2}$$

over $x \in D$, where $D = \{x | x_1 + x_2 > 1\}$. Note that C is the closure of D and $\beta = \inf\{f(x) | x \in D\}$. Setting the partial derivatives to zero, we find that

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

As $k \to +\infty$ the sequence $\{x^k\}$ converges to $(\frac{1}{2}, \frac{1}{2})$, which is the solution.

In this example the auxiliary function $-\log(x_1 + x_2 - 1)$ serves to limit the calculations to those x satisfying the constraint $x_1 + x_2 > 1$, while also permitting us to get x^k in closed form. The minus sign may seem unnecessary, since $\log(x_1 + x_2 - 1)$ would similarly restrict the x. However, without the minus sign there is no minimizer of $B_k(x)$ within D.

1.2.2 Barrier-Function Methods: Another Example

Now we want to minimize the function $f(x) = x_1 + x_2$, subject to the constraints $-x_1^2 + x_2 \geq 0$ and $x_1 \geq 0$. Once again, we use the log function to create our barrier function. We minimize

$$B_k(x) = f(x) + \frac{1}{k} b(x) = x_1 + x_2 + \frac{1}{k}(-\log(-x_1^2 + x_2) - \log x_1) \tag{1.3}$$

to get x^k. With a bit of algebra we find that

$$x_1^k = -\frac{1}{4} + \frac{1}{4}\sqrt{1 + \frac{8}{k}},$$

and

$$x_2^k = \frac{1}{16}\left(-1 + \sqrt{1 + \frac{8}{k}}\right)^2 + \frac{1}{k}.$$

As $k \to +\infty$ the sequence $\{x^k\}$ converges to $(0, 0)$, which is the answer.

1.2.3 Barrier-Function Methods for the Basic Problem

Now the problem is to minimize $f : X \to \mathbb{R}$, subject to $x \in C$. We select $b : X \to (-\infty, +\infty]$ with $C = \{x | b(x) < +\infty\}$. For each k we minimize $B_k(x) = f(x) + \frac{1}{k}b(x)$ over all $x \in X$ to get x^k, which must necessarily lie in C. Formulated this way, the method is not yet in AF form. Nevertheless, we have the following proposition.

Proposition 1.2 *The sequence $\{b(x^k)\}$ is nondecreasing, and the sequence $\{f(x^k)\}$ is nonincreasing and converges to $\beta = \inf_{x \in C} f(x)$.*

Proof: From $B_k(x^{k-1}) \geq B_k(x^k)$ and $B_{k-1}(x^k) \geq B_{k-1}(x^{k-1})$, for $k = 2, 3, ...$, it follows easily that

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

Suppose that $\{f(x^k)\} \downarrow \beta^* > \beta$. Then there is $z \in C$ with

$$f(x^k) \geq \beta^* > f(z) \geq \beta,$$

for all k. Then

$$\frac{1}{k}(b(z) - b(x^k)) \geq f(x^k) - f(z) \geq \beta^* - f(z) > 0,$$

for all k. But the sequence $\{\frac{1}{k}(b(z) - b(x^k))\}$ converges to zero, which contradicts the assumption that $\beta^* > \beta$. ∎

The proof of Proposition 1.2 depended heavily on the details of the barrier-function method. Now we reformulate the barrier-function method as an AF method, and obtain a different proof of Proposition 1.2 that leads to the definition of the SUMMA class of of AF methods [47, 55].

Minimizing $B_k(x) = f(x) + \frac{1}{k}b(x)$ to get x^k is equivalent to minimizing $kf(x) + b(x)$, which, in turn, is equivalent to minimizing

$$G_k(x) = f(x) + g_k(x),$$

where

$$g_k(x) = [(k-1)f(x) + b(x)] - [(k-1)f(x^{k-1}) + b(x^{k-1})].$$

Clearly, $g_k(x) \geq 0$ and $g_k(x^{k-1}) = 0$. Now we have the AF form of the method. Here is a different proof of Proposition 1.2.

A simple calculation shows that

$$G_k(x) - G_k(x^k) = g_{k+1}(x), \tag{1.4}$$

for all $x \in X$. Suppose that the nonincreasing sequence $\{f(x^k)\}$ converges to some $\beta^* > \beta$. Then there is $z \in C$ with $\beta^* > f(z) \geq \beta$. Then

$$g_k(z) - g_{k+1}(z) = g_k(z) - G_k(z) + G_k(x^k)$$

$$= g_k(z) - f(z) - g_k(z) + f(x^k) + g_k(x^k) \geq \beta^* - f(z) > 0.$$

This is impossible, since $\{g_k z)$ is then a nonincreasing sequence of nonnegative numbers whose successive differences are bounded below by the positive number $\beta^* - f(z)$.

1.2.4 The SUMMA Class

Close inspection of the second proof of Proposition 1.2 reveals that Equation (1.4) is unnecessary; all we need is the SUMMA Inequality

$$G_k(x) - G_k(x^k) \geq g_{k+1}(x), \tag{1.5}$$

for all $x \in X$. All iterative methods that can be reformulated as AF methods for which the inequality in (1.5) holds are said to belong to the SUMMA class of methods. This may seem to be a quite restricted class of methods, but, as we shall see, that is far from the case. Many well known iterative methods fall into the SUMMA class [47, 55].

1.2.5 Cross-Entropy Methods

For $a > 0$ and $b > 0$, let the cross-entropy or Kullback-Leibler (KL) distance [117] from a to b be

$$KL(a, b) = a \log \frac{a}{b} + b - a, \tag{1.6}$$

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

$$KL(x, z) = \sum_{j=1}^{J} KL(x_j, z_j). \tag{1.7}$$

Then $KL(x,z) \geq 0$ and $KL(x,z) = 0$ if and only if $x = z$. Unlike the Euclidean distance, the KL distance is not symmetric; $KL(x,y)$ and $KL(y,x)$ are distinct. We can obtain different approximate solutions of a nonnegative system of linear equations $Px = y$ by minimizing $KL(Px,y)$ and $KL(y,Px)$ with respect to nonnegative x. The SMART minimizes $KL(Px,y)$, while the EMML algorithm minimizes $KL(y,Px)$. Both are iterative algorithms in the SUMMA class, and are best developed using the *alternating minimization* (AM) framework.

1.2.6 Alternating Minimization

Let $\Theta : P \times Q \to (-\infty, +\infty]$, where P and Q are arbitrary nonempty sets. In the *alternating minimization* (AM) method we minimize $\Theta(p, q^{n-1})$ over $p \in P$ to get p^n and then minimize $\Theta(p^n, q)$ over $q \in Q$ to get q^n. We want

$$\{\Theta(p^n, q^n)\} \downarrow \beta = \inf\{\Theta(p,q) | p \in P, q \in Q\}. \tag{1.8}$$

In [81] Csiszár and Tusnády show that if the function Θ possesses what they call the *five-point property*, then Equation (1.8) holds. There seemed to be no convincing explanation of why the five-point property should be used, except that it works. I was quite surprised when I discovered that the AM method can be reformulated as an AF method to minimize a function of the single variable p and that the five-point property becomes precisely the SUMMA condition.

1.2.7 Penalty-Function Methods

Once again, we want to minimize $f : X \to \mathbb{R}$, subject to $x \in C$. We select a penalty function $p : X \to [0, +\infty)$ with $p(x) = 0$ if and only if $x \in C$. Then, for each k, we minimize

$$P_k(x) = f(x) + kp(x),$$

over all x, to get x^k. Here is a simple example of the use of penalty-function methods.

Let us minimize the function $f(x) = (x+1)^2$, subject to $x \geq 0$. We let $p(x) = 0$ for $x \geq 0$, and $p(x) = x^2$, for $x < 0$. Then $x^k = -\frac{1}{k+1}$, which converges to zero, the correct answer, as $k \to +\infty$. Note that x^k is not in $C = \mathbb{R}_+$, which is why such methods are called *exterior-point methods*.

Clearly, it is equivalent to minimize

$$p(x) + \frac{1}{k} f(x),$$

which gives the penalty-function method the form of a barrier-function

method. From Proposition 1.2 it follows that the sequence $\{p(x^k)\}$ is non-increasing and converges to zero, while the sequence $\{f(x^k)\}$ is nondecreasing, and, as we can easily show, converges to some $\gamma \leq \beta$.

Without imposing further structure on X and f we cannot conclude that $\{f(x^k)\}$ converges to β. The reason is that, in the absence of further structure, such as the continuity of f, what f does within C is unrelated to what it does outside C. If, for some f, we do have $\{f(x^k)\}$ converging to β, we can replace $f(x)$ with $f(x) - 1$ for x not in C, while leaving $f(x)$ unchanged for x in C. Then β remains unaltered, while the new sequence $\{f(x^k)\}$ converges to $\gamma = \beta - 1$.

1.3 Fixed-Point Methods

We turn now to fixed-point (FP) methods, the second class of methods we shall discuss in these notes. For our discussion of fixed-point methods we shall impose some structure on X, although, as we shall see, it may not be that of a metric space. However, most of the examples we shall present will be in the setting of a Hilbert space $\mathcal{H}$, usually $\mathcal{H} = \mathbb{R}^J$.

When we use an FP method to solve a problem iteratively, we select an operator $T : X \to X$ such that z solves the problem if and only if z is a fixed point of T, that is, $Tz = z$. The set of fixed points of T will be denoted $\text{Fix}(T)$. Our iterative method is then to let $x^k = Tx^{k-1}$, for $k = 1, 2, ...$ If we are trying to minimize $f : X \to \mathbb{R}$, subject to $x \in C$, then, as before, we want the sequence $\{f(x^k)\}$ to converge to β. If we have imposed some topology on X, we can also ask for the sequence $\{x^k\}$ to converge, at least weakly, to a solution of the problem.

Definition 1.1 *An operator $T : \mathcal{H} \to \mathcal{H}$ is* convergent *if T has at least one fixed point and the sequence $\{T^k x^0\}$ converges weakly to a fixed point of T, for every starting point x^0.*

1.3.1 Gradient Descent Algorithms

Suppose that we want to minimize $f : \mathcal{H} \to \mathbb{R}$. When f is Gâteaux differentiable the derivative of f at x is the gradient, $\nabla f(x)$. A gradient-descent algorithm has the iterative step

$$x^k = x^{k-1} - \gamma_k \nabla f(x^{k-1}), \tag{1.9}$$

where the step-length parameters γ_k are adjusted at each step. When f is Gâteaux differentiable at x, the one-sided directional derivative of f at x

and in the direction d, denoted $f'_+(x;d)$, is given by

$$f'_+(x;d) = \langle \nabla f(x), d \rangle.$$

When f is convex and Gâteaux differentiable and the gradient ∇f is L-Lipschitz continuous, that is,

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

for all x and y, we can modify the iteration in Equation (1.9). If $0 < \gamma < \frac{2}{L}$, and

$$x^k = x^{k-1} - \gamma \nabla f(x^{k-1}), \tag{1.10}$$

then the sequence $\{x^k\}$ converges to a zero of ∇f, whenever there are such zeros. Now that γ is independent of k, we can take $T = I - \gamma \nabla f$, with I the identity operator, and write the iteration as $x^k = Tx^{k-1}$. As we shall show later, the operator T is convergent. A point z is a fixed point of T if and only if $\nabla f(z) = 0$, and so if and only if z minimizes f.

Definition 1.2 *For any function $f : \mathcal{H} \to \mathbb{R}$ the* epigraph *of f is the set*

$$\text{epi}(f) = \{(x, \gamma) \in \mathcal{H} \times \mathbb{R} | f(x) \leq \gamma\}.$$

Definition 1.3 *A function $f : \mathcal{H} \to \mathbb{R}$ is* lower semi-continuous *or* closed *if* $\text{epi}(f)$ *is closed in $\mathcal{H} \times \mathbb{R}$.*

If $f : \mathcal{H} \to \mathbb{R}$ is closed and convex, then $\text{epi}(f)$ is a nonempty, closed, and convex set.

Note that when $f : \mathbb{R}^J \to \mathbb{R}$ is convex, it is continuous. It is also true in the infinite-dimensional case, provided that f is closed as well ([15], Corollary 8.30). For any convex $f : \mathcal{H} \to \mathbb{R}$, Gâteaux differentiability and Fréchet differentiability are equivalent for finite-dimensional $\mathcal{H}$, but not necessarily equivalent in the case of infinite-dimensional Hilbert space. We shall use the word "differentiable" to mean Gâteaux differentiable. Whenever f is differentiable and ∇f is continuous, f is Fréchet differentiable. Therefore, if ∇f is L-Lipschitz continuous, then f is Fréchet differentiable.

A function $f : \mathcal{H} \to [-\infty, +\infty]$ is *proper* if there is no x with $f(x) = -\infty$ and some x with $f(x) < +\infty$. All the functions we consider are proper.

1.3.2 Projected Gradient Descent

For any nonempty, closed convex subset C of a Hilbert space $\mathcal{H}$ and any x in $\mathcal{H}$ there is a unique $z \in C$ closest to x. This z is denoted $z = P_C x$ and the operator P_C is called an *orthogonal projection* (or *metric projection*) operator.

Suppose now that we want to minimize a differentiable closed convex function $f : \mathcal{H} \to \mathbb{R}$ over $x \in C$, where C is a nonempty closed, convex subset of $\mathcal{H}$. Assume that ∇f is L-Lipschitz continuous, and $0 < \gamma < \frac{2}{L}$. The *projected gradient descent* algorithm has the iterative step

$$x^k = P_C(x^{k-1} - \gamma \nabla f(x^{k-1})). \tag{1.11}$$

The sequence $\{x^k\}$ converges weakly to a point $z \in C$ with $f(z) \leq f(x)$, for all $x \in C$, whenever such a point z exists. It is not hard to show that such z are the fixed points of the operator $T = P_C(I - \gamma \nabla f)$, which is a convergent operator.

1.3.3 Solving $Ax = b$

Suppose that A is a real M by N matrix and b is a member of $\mathbb{R}^M$. We want $x \in \mathbb{R}^N$ so that $Ax = b$, or, if there are no such x, then we want an x that minimizes the function $f(x) = \frac{1}{2}\|Ax - b\|^2$, where the norm is the two-norm (that is, the usual Euclidean norm). There is a closed-form solution for this problem: $x = (A^T A)^{-1} A^T b$, provided that $A^T A$ is invertible. In many applications in image processing and remote sensing the matrix A can be quite large, with M and N in the tens of thousands. In such cases, using the closed-form expression for the solution is not practical and we turn to iterative methods.

The function $f(x) = \frac{1}{2}\|Ax - b\|^2$ is differentiable and its derivative,

$$\nabla f(x) = A^T(Ax - b),$$

is L-Lipschitz continuous for $L = \rho(A^T A)$, the spectral radius of $A^T A$, which, in this case, is the largest eigenvalue of $A^T A$. Applying the gradient descent algorithm in the previous section, we get the Landweber iteration [118, 20],

$$x^k = x^{k-1} - \gamma A^T(Ax^{k-1} - b), \tag{1.12}$$

for $0 < \gamma < \frac{2}{\rho(A^T A)}$. The sequence $\{x^k\}$ converges to the minimizer of $f(x)$ closest to x^0.

1.3.4 Projected Landweber Algorithm

Suppose that we want to find $x \in C \subseteq \mathbb{R}^N$ such that $Ax = b$, or, failing that, to minimize the function $\frac{1}{2}\|Ax - b\|^2$ over $x \in C$. Applying the projected gradient descent algorithm, we get the *projected Landweber* algorithm [20],

$$x^k = P_C(x^{k-1} - \gamma A^T(Ax^{k-1} - b)). \tag{1.13}$$

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

1.3.5 The Split Feasibility Problem

Let C and Q be nonempty closed, convex subsets of $\mathbb{R}^N$ and $\mathbb{R}^M$, respectively, and A a real M by N matrix. The *split feasibility problem* (SFP) is to find $x \in C$ with $Ax \in Q$, or, failing that, to minimize the function $f(x) = \frac{1}{2}\|P_Q Ax - Ax\|^2$ over $x \in C$. It can be shown [55] that $f(x)$ is differentiable and its gradient is

$$\nabla f(x) = A^T(I - P_Q)Ax.$$

The gradient is again L-Lipschitz continuous for $L = \rho(A^T A)$. Applying the gradient descent algorithm we have the CQ algorithm [42, 43]:

$$x^k = P_C(x^{k-1} - \gamma A^T(I - P_Q)Ax^{k-1}). \tag{1.14}$$

Solutions of the problem are the fixed points of the convergent operator $T : \mathcal{H} \to \mathcal{H}$ given by $T = P_C(I - \gamma A^T(I - P_Q)A)$.

In [66, 62] Yair Censor and his colleagues modified the CQ algorithm and used their modified method to derive protocols for intensity modified radiation therapy (IMRT).

1.3.6 Firmly Nonexpansive Operators

We are interested in operators T that are convergent. For such operators we often find that $\|x^{k+1} - x^k\| \leq \|x^k - x^{k-1}\|$ for each k. This leads us to the definition of *nonexpansive* operators.

Definition 1.4 *An operator T on $\mathcal{H}$ is* nonexpansive (ne) *if, for all x and y, we have*

$$\|Tx - Ty\| \leq \|x - y\|.$$

Nonexpansive operators need not be convergent, as the ne operator $T = -I$ illustrates.

As we shall see later, the operators $T = P_C$ are nonexpansive. In fact, the operators P_C have a much stronger property; they are firmly nonexpansive.

Definition 1.5 *An operator T on $\mathcal{H}$ is* firmly nonexpansive (fne) *if, for every x and y, we have*

$$\langle Tx - Ty, x - y\rangle \geq \|Tx - Ty\|^2.$$

If T is fne then T is convergent. The class of fne operators is smaller than the class of ne operators and does yield convergent iterative sequences. However, the product or composition of two or more fne operators need not be fne, which limits the usefulness of this class of operators. Even the product of P_{C_1} and P_{C_2} need not be fne. We need to find a class of convergent operators that is closed to finite products.

1.3.7 Averaged Operators

It can be shown easily that an operator F is fne if and only if there is a nonexpansive operator N such that

$$F = \frac{1}{2}I + \frac{1}{2}N.$$

Definition 1.6 *An operator $A : \mathcal{H} \to \mathcal{H}$ is* α-averaged (α-av) *if there is a nonexpansive operator N such that*

$$A = (1-\alpha)I + \alpha N,$$

for some α in the interval $(0, 1)$. If A is α-av *for some α then A is an* averaged (av) *operator.*

All averaged operators are nonexpansive, all firmly nonexpansive operators are averaged, the class of averaged operators is closed to finite products, and averaged operators are convergent. In other words, the class of averaged operators is precisely the class that we are looking for.

1.3.8 Useful Properties of Operators on $\mathcal{H}$

It turns out that properties of an operator T are often more easily studied in terms of properties of its complement, $G = I - T$. The following two identities are easy to prove and are quite helpful. For any operator $T : \mathcal{H} \to \mathcal{H}$ and $G = I - T$ we have

$$\|x-y\|^2 - \|Tx-Ty\|^2 = 2\langle Gx-Gy, x-y\rangle - \|Gx-Gy\|^2, \quad (1.15)$$

and

$$\langle Tx-Ty, x-y\rangle - \|Tx-Ty\|^2 = \langle Gx-Gy, x-y\rangle - \|Gx-Gy\|^2. \quad (1.16)$$

Definition 1.7 *An operator $G : \mathcal{H} \to \mathcal{H}$ is* ν-inverse strongly monotone (ν-ism) *for some $\nu > 0$ if*

$$\langle Gx-Gy, x-y\rangle \geq \nu\|Gx-Gy\|^2,$$

for all x and y.

Clearly, if G is ν-ism then γG is $\frac{\nu}{\gamma}$-ism. Using the two identities in (1.15) and (1.16) it is easy to prove the following theorem.

Theorem 1.1 *Let $T : \mathcal{H} \to \mathcal{H}$ be arbitrary and $G = I - T$. Then*

1. *T is ne if and only if G is ν-ism for $\nu = \frac{1}{2}$;*
2. *T is α-av if and only if G is ν-ism for $\nu = \frac{1}{2\alpha}$, for some $0 < \alpha < 1$;*

3. *T is fne if and only if G is ν-ism for $\nu = 1$.*

4. *T is fne if and only if G is fne;*

5. *If G is ν-ism and $0 < \mu \leq \nu$, then G is μ-ism.*

1.3.9 Subdifferentials and Subgradients

Subdifferentials and subgradients are important tools in optimization, particularly in convex optimization.

Definition 1.8 *Let $f : \mathcal{H} \to [-\infty, \infty]$. Then $\partial f(x)$ is the* subdifferential *of f at x, defined by*

$$\partial f(x) = \{u | \langle u, z - x \rangle + f(x) \leq f(z)\}, \tag{1.17}$$

for all z. The members of $\partial f(x)$ are the subgradients *of f at x.*

It is easy to see that, if $z \in \mathcal{H}$ minimizes $f(x)$ over $x \in \mathcal{H}$, then $0 \in \partial f(z)$.

There is a subtle point to be aware of here: if f is differentiable, but not convex, $\nabla f(x)$ need not be a member of $\partial f(x)$, which may be empty. If f is closed and convex and $\partial f(x)$ is a singleton set, then f is differentiable and $\partial f(x) = \{\nabla f(x)\}$. The following proposition provides a characterization of convexity for nondifferentiable functions.

Proposition 1.3 *A function $f : \mathcal{H} \to \mathbb{R}$ is closed and convex if and only if $\partial f(x)$ is nonempty, for every $x \in \mathcal{H}$.*

Proof: When f is closed and convex, its epigraph is a closed convex set. We can then use orthogonal projection to find a supporting hyperplane for the epigraph at the point $(x, f(x))$. From the normal to the hyperplane we can construct a member of $\partial f(x)$ [56]. Now we prove the converse.

By Proposition 17.39 of [15], if $f : \mathcal{H} \to \mathbb{R}$ is convex and $\partial f(x)$ is nonempty, for each x, then f is closed. Now let x and y be arbitrary in $\mathcal{H}$, $z = (1 - \alpha)x + \alpha y$, for some $\alpha \in (0, 1)$, and u a member of $\partial f(z)$. Then

$$f(x) - f(z) \geq \langle u, x - z \rangle = \alpha \langle u, x - y \rangle,$$

and

$$f(y) - f(z) \geq \langle u, y - z \rangle = (1 - \alpha)\langle u, y - x \rangle = -(1 - \alpha)\langle u, x - y \rangle.$$

Therefore,

$$(1 - \alpha)(f(x) - f(z)) \geq (1 - \alpha)\alpha \langle u, x - y \rangle \geq \alpha(f(z) - f(y)),$$

and so

$$(1 - \alpha)f(x) + \alpha f(y) \geq (1 - \alpha)f(z) + \alpha f(z) = f(z).$$

∎

Proposition 1.4 *For any $f : \mathcal{H} \to \mathbb{R}$ and $g : \mathcal{H} \to \mathbb{R}$ we have*

$$\partial f(x) + \partial g(x) \subseteq \partial(f+g)(x). \tag{1.18}$$

If f and g are closed and convex, then

$$\partial f(x) + \partial g(x) = \partial(f+g)(x). \tag{1.19}$$

Proof: The containment in (1.18) follows immediately from the definition of the subdifferential. For the proof of (1.19) see Corollary 16.38 of [15]. ∎

In some discussions, convex functions may be allowed to take on the value $+\infty$. In such cases only the containment in (1.18) may hold; see Corollary 16.38 of [15].

Corollary 1.1 *If $f : \mathcal{H} \to \mathbb{R}$ and $g : \mathcal{H} \to \mathbb{R}$ are both closed and convex, and $f + g = h$ is differentiable, then both f and g are differentiable.*

Proof: From Proposition 1.4 we have

$$\partial f(x) + \partial g(x) \subseteq \partial(f+g)(x) = \partial h(x) = \{\nabla h(x)\}.$$

Since both $\partial f(x)$ and $\partial g(x)$ are nonempty, they must be singleton sets. Therefore, both functions are differentiable, according to Proposition 17.26 of [15]. ∎

1.3.10 Monotone Operators

There is an interesting connection between fne operators and monotone operators.

Definition 1.9 *A set-valued function $B : \mathcal{H} \to 2^{\mathcal{H}}$ is said to be* monotone *if, for each x and y in $\mathcal{H}$ and $u \in B(x)$ and $v \in B(y)$ we have $\langle u-v, x-y\rangle \geq 0$. If there is no monotone $A : \mathcal{H} \to 2^{\mathcal{H}}$ with $B(x) \subseteq A(x)$ for all x, then B is a* maximal *monotone operator.*

If $f : \mathcal{H} \to \mathbb{R}$ is closed and convex then $B(x) = \partial f(x)$ defines a monotone operator. In particular, if f is also differentiable then $T = \nabla f$ is a monotone operator. We have the following proposition.

Proposition 1.5 *If B is monotone and $x \in z + B(z)$ and $x \in y + B(y)$, then $z = y$*

The proof is not difficult and we leave it to the reader.

It follows from Proposition 1.5 that z is uniquely defined by the inclusion $x \in z + B(z)$ and we write $z = J_B x$. The operator $J_B : \mathcal{H} \to \mathcal{H}$ is the *resolvent* of the monotone operator B. Sometimes we write $J_B = (I+B)^{-1}$. The operator J_B is fne and T is a fne operator if and only if there is monotone operator $B : \mathcal{H} \to 2^{\mathcal{H}}$ such that $T = J_B$. Given operator T, define $B(x) = T^{-1}(\{x\}) - x$. Then T is fne if and only if B is monotone.

1.3.11 The Baillon–Haddad Theorem

The Baillon–Haddad Theorem [4, 14] provides one of the most important links between fixed-point methods and iterative optimization. The proof we give here is new [52]. It is the first elementary proof of this theorem and depends only on basic properties of convex functions. The non-elementary proof of this theorem in [100] was repeated in the book [46]. The proof given here and in [52] is closely related to that given in the book [55].

Definition 1.10 *Let $f : \mathcal{H} \to \mathbb{R}$ be convex and differentiable. The* Bregman distance *associated with f is $D_f(x, y)$ given by*

$$D_f(x, y) = f(x) - f(y) - \langle \nabla f(y), x - y \rangle.$$

Then $D_f(x, y) \geq 0$, and $D_f(x, x) = 0$. If f is strictly convex, then $D_f(x, y) = 0$ if and only if $x = y$.

Theorem 1.2 *Let $f : \mathcal{H} \to \mathbb{R}$ be convex and differentiable, and let $q(x) = \frac{1}{2}\|x\|^2$. The following are equivalent:*

1. *$g = q - f$ is convex;*
2. *$\frac{1}{2}\|z - x\|^2 \geq D_f(z, x)$ for all x and z;*
3. *$T = \nabla f$ is firmly nonexpansive;*
4. *$T = \nabla f$ is nonexpansive and f is Fréchet differentiable.*

Proof:

- (1. implies 2.) Because g is convex, we have

  $$g(z) \geq g(x) + \langle \nabla g(x), z - x \rangle,$$

  which is easily shown to be equivalent to

  $$\frac{1}{2}\|z - x\|^2 \geq f(z) - f(x) - \langle \nabla f(x), z - x \rangle = D_f(z, x).$$

- (2. implies 3.) Fix y and define $d(x)$ by

  $$d(x) = D_f(x, y) = f(x) - f(y) - \langle \nabla f(y), x - y \rangle \geq 0.$$

  Then

  $$\nabla d(x) = \nabla f(x) - \nabla f(y)$$

  and $D_f(z, x) = D_d(z, x)$ for all z and x. Therefore, we have

  $$\frac{1}{2}\|z - x\|^2 \geq D_d(z, x) = d(z) - d(x) - \langle \nabla d(x), z - x \rangle.$$

Now let $z - x = \nabla f(y) - \nabla f(x)$, so that

$$d(x) = D_f(x, y) \geq \frac{1}{2}\|\nabla f(x) - \nabla f(y)\|^2.$$

Similarly,

$$D_f(y, x) \geq \frac{1}{2}\|\nabla f(x) - \nabla f(y)\|^2.$$

Adding these two inequalities gives

$$\langle \nabla f(x) - \nabla f(y), x - y\rangle \geq \|\nabla f(x) - \nabla f(y)\|^2.$$

- (3. implies 4.) Clearly, if ∇f is firmly nonexpansive, it is also nonexpansive. Since it is then continuous, f must be Fréchet differentiable.
- (4. implies 1.) From $\nabla g(x) = x - \nabla f(x)$ we get

$$\langle \nabla g(x) - \nabla g(y), x - y\rangle = \|x - y\|^2 - \langle \nabla f(x) - \nabla f(y), x - y\rangle$$

$$\geq \|x - y\|(\|x - y\| - \|\nabla f(x) - \nabla f(y)\|) \geq 0.$$

Therefore, g is convex.

∎

We get a slightly more general version of Theorem 4.1, but with a slightly less elementary proof, if we assume that f is closed and omit the assumption that f be differentiable. Once we assume 1., the differentiablility of f follows from Proposition 1.4 and Corollary 1.1.

As was mentioned previously, the Baillon–Haddad Theorem plays an important role in linking fixed-point algorithms to optimization. Suppose that $f : \mathcal{H} \to \mathbb{R}$ is convex and differentiable, and its gradient, ∇f, is L-Lipschitz continuous. Then the gradient of the function $g = \frac{1}{L}f$ is ne, and so ∇g is fne. As we shall see in Chapter 4, it follows from the theory of averaged operators and the Krasnosel'skii-Mann-Opial Theorem 4.2 that the operator $I - \gamma\nabla f$ is an averaged operator, therefore a convergent operator, for $0 < \gamma < \frac{2}{L}$.

In [14] Bauschke and Combettes extend the Baillon–Haddad Theorem to include several other equivalent conditions. These additional conditions involve definitions and results that are not elementary; we shall return to their expanded version of the theorem in Chapter 7.

Chapter 2

Auxiliary-Function Methods and Examples

2.1 Auxiliary-Function Methods

We suppose that $f : X \to \mathbb{R}$ and $C \subseteq X$, where X is an arbitrary nonempty set. An iterative algorithm is in the AF class if, for $k = 1, 2, ...$ we minimize $G_k(x) = f(x) + g_k(x)$ over $x \in X$ to get x^k, where $g_k(x) \geq 0$ and $g_k(x^{k-1}) = 0$. As we saw previously, the sequence $\{f(x^k)\}$ is then nonincreasing. We want the sequence $\{f(x^k)\}$ to converge to $b = \inf\{f(x)|x \in C\}$. If C is a proper subset of X we replace f with $f + \iota_C$ at the beginning. In that case every x^k lies in C.

2.2 Majorization Minimization

Majorization minimization (MM), also called *optimization transfer*, is a technique used in statistics to convert a difficult optimization problem into a sequence of simpler ones [138, 18, 121]. The MM method requires that we majorize the objective function $f(x)$ with $g(x|y)$, such that $g(x|y) \geq f(x)$, for all x and y, and $g(y|y) = f(y)$. At the kth step of the iterative algorithm we minimize the function $g(x|x^{k-1})$ to get x^k.

The MM methods are members of the AF class. At the kth step of an MM iteration we minimize

$$G_k(x) = f(x) + [g(x|x^{k-1}) - f(x)] = f(x) + d(x, x^{k-1}), \tag{2.1}$$

where $d(x, z) = g(x|z) - f(x)$ is a distance function satisfying $d(x, z) \geq 0$

and $d(z,z) = 0$. Since $g_k(x) = d(x, x^{k-1}) \geq 0$ and $g_k(x^{k-1}) = 0$, MM methods are also AF methods; it then follows that the sequence $\{f(x^k)\}$ is nonincreasing.

All MM algorithms have the form $x^k = Tx^{k-1}$, where T is the operator defined by

$$Tz = \text{argmin}_x \{f(x) + d(x,z)\}. \tag{2.2}$$

If $d(x,z) = \frac{1}{2}\|x-z\|_2^2$, then T is Moreau's proximity operator $Tz = \text{prox}_f(z)$ [131, 132, 133], which we shall discuss in some detail later.

2.3 The Method of Auslander and Teboulle

The method of Auslander and Teboulle [2] is a particular example of an MM algorithm. We take C to be a closed, nonempty, convex subset of $\mathbb{R}^J$, with interior U. At the kth step of their method one minimizes a function

$$G_k(x) = f(x) + d(x, x^{k-1}) \tag{2.3}$$

to get x^k. Their distance $d(x,y)$ is defined for x and y in U, and the gradient with respect to the first variable, denoted $\nabla_1 d(x,y)$, is assumed to exist. The 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 U, with $H(a,a) = 0$ and

$$\langle \nabla_1 d(b,a), c-b \rangle \leq H(c,a) - H(c,b), \tag{2.4}$$

for all c in U.

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

$$\langle \nabla_1 d(b,a), c-b \rangle = D_h(c,a) - D_h(c,b) - D_h(b,a) \tag{2.5}$$

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 [2].

The method of Auslander and Teboulle seems not to be a particular case of SUMMA. However, we can adapt the proof of Proposition 1.2 to prove the analogous result for their method. We assume that $f(\hat{x}) \leq f(x)$, for all x in C.

Theorem 2.1 *For $k = 2, 3, ...$, let x^k minimize the function*

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

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

Proof: We know 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 U with

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

for all k. Since x^k minimizes $G_k(x)$, 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.

2.4 The EM Algorithm

The *expectation maximization maximum likelihood* (EM) "algorithm" is not a single algorithm, but a framework, or, as the authors of [18] put it, a "prescription" , for constructing algorithms. Nevertheless, we shall refer to it as the EM algorithm.

The EM algorithm is always presented within the context of statistical likelihood maximization, but the essence of this method is not stochastic; the EM algorithms can be shown to form a subclass of AF methods. We

present now the essential aspects of the EM algorithm without relying on statistical concepts.

The problem is to maximize a nonnegative function $f : Z \to \mathbb{R}$, where Z is an arbitrary set. In the stochastic context $f(z)$ is a likelihood function of the parameter vector z. We assume that there is $z^* \in Z$ with $f(z^*) \geq f(z)$, for all $z \in Z$.

We also assume that there is a nonnegative function $h : \mathbb{R}^J \times Z \to \mathbb{R}$ such that

$$f(z) = \int h(x, z)dx.$$

Having found z^{k-1}, we maximize the function

$$H(z^{k-1}, z) = \int h(x, z^{k-1}) \log h(x, z)dx \tag{2.6}$$

to get z^k. Adopting such an iterative approach presupposes that maximizing $H(z^{k-1}, z)$ is simpler than maximizing $f(z)$ itself. This is the case with the EM algorithms.

One of the most useful and easily proved facts about the Kullback-Leibler distance is contained in the following lemma.

Lemma 2.1 *For nonnegative vectors x and z, with $z_+ = \sum_{j=1}^J z_j > 0$, we have*

$$KL(x, z) = KL(x_+, z_+) + KL(x, \frac{x_+}{z_+}z). \tag{2.7}$$

This lemma can be extended by replacing the summation with integration over the variable x. Then we obtain the following useful identity; we simplify the notation by setting $h(z) = h(x, z)$.

Lemma 2.2 *For $f(z)$ and $h(x, z)$ as above, and z and w in Z, with $f(w) > 0$, we have*

$$KL(h(z), h(w)) = KL(f(z), f(w)) + KL(h(z), (f(z)/f(w))h(w)). \tag{2.8}$$

Maximizing $H(z^{k-1}, z)$ is equivalent to minimizing

$$G_k(z) = G(z^{k-1}, z) = -f(z) + KL(h(z^{k-1}), h(z)), \tag{2.9}$$

where

$$g_k(z) = KL(h(z^{k-1}), h(z)) = \int KL(h(x, z^{k-1}), h(x, z))dx. \tag{2.10}$$

Since $g_k(z) \geq 0$ for all z and $g_k(z^{k-1}) = 0$, we have an AF method. Without additional restrictions, we cannot conclude that $\{f(z^k)\}$ converges to $f(z^*)$.

We get z^k by minimizing $G_k(z) = G(z^{k-1}, z)$. When we minimize $G(z, z^k)$, we get z^k again. Therefore, we can put the EM algorithm into the alternating minimization (AM) framework of Csiszár and Tusnády [81], to be discussed in Chapter 6.

Chapter 3

The SUMMA Class

3.1 The SUMMA Class of Algorithms

Through our examination of barrier-function methods we discovered the SUMMA condition [47, 55]:

$$G_k(x) - G_k(x^k) \geq g_{k+1}(x), \tag{3.1}$$

for all $x \in X$. The SUMMA condition gives $\{f(x^k)\} \downarrow \beta = \inf_x f(x)$. As we saw, barrier-function methods can be reformulated as SUMMA algorithms and penalty-function methods can be reformulated as barrier-function methods. Although the SUMMA condition may seem quite restrictive, the class of SUMMA algorithms is extensive. In this chapter we examine several of these algorithms.

3.2 Proximal Minimization

Let $h : \mathcal{H} \to (-\infty, \infty]$ be convex and differentiable on the interior of dom $h = \{x | h(x) \in \mathbb{R}\}$, its effective domain. The *Bregman distance*

associated with the function h is

$$D_h(x,y) = h(x) - h(y) - \langle \nabla h(y), x - y \rangle. \tag{3.2}$$

Then $D_h(x,y) \geq 0$, $D_h(x,x) = 0$, and if h is strictly convex then $D_h(x,y) = 0$ if and only if $x = y$.

Let $f : \mathcal{H} \to (-\infty, +\infty]$ be a closed convex function. Let $h : \mathcal{H} \to (-\infty, \infty]$ be another 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 $\hat{x}$. Our objective is to minimize $f(x)$ over x in $C = \overline{D}$.

3.2.1 The PMA

At the kth step of a *proximal minimization algorithm* (PMA) [72, 40], we minimize the function

$$G_k(x) = f(x) + D_h(x, x^{k-1}), \tag{3.3}$$

to get x^k. The Bregman distance D_h is sometimes called a *proximity function*. The function

$$g_k(x) = D_h(x, x^{k-1}) \tag{3.4}$$

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

$$G_k(x) - G_k(x^k) \geq D_h(x, x^k) = g_{k+1}(x) \geq 0, \tag{3.5}$$

so any PMA is in the SUMMA class.

The Newton-Raphson algorithm for minimizing a twice differentiable function $f : \mathbb{R}^J \to \mathbb{R}$ has the iterative step

$$x^k = x^{k-1} - \nabla^2 f(x^{k-1})^{-1} \nabla f(x^{k-1}). \tag{3.6}$$

Suppose now that f is also convex. It is interesting to note that, having calculated x^{k-1}, we can obtain x^k by minimizing

$$G_k(x) = f(x) + (x - x^{k-1})^T \nabla^2 f(x^{k-1})(x - x^{k-1}) - D_f(x, x^{k-1}). \tag{3.7}$$

3.2.2 Difficulties with the PMA

The PMA can present some computational obstacles. When we minimize $G_k(x)$ to get x^k we find that we must solve the equation

$$\nabla h(x^{k-1}) - \nabla h(x^k) \in \partial f(x^k), \tag{3.8}$$

where the set $\partial f(x)$ is the subdifferential of f at x. When $f(x)$ is differentiable $\partial f(x) = \{\nabla f(x)\}$ and we must solve

$$\nabla f(x^k) + \nabla h(x^k) = \nabla h(x^{k-1}). \tag{3.9}$$

A particular case of the PMA, called the IPA for *interior-point algorithm* [40, 47], is designed to overcome these computational obstacles. We discuss the IPA later in this chapter. Another modification of the PMA that is similar to the IPA is the *forward-backward splitting* (FBS) method, to be discussed in Chapter 4.

3.2.3 All PMA are in the SUMMA Class

We show now that all PMA are in the SUMMA class. We remind the reader that $f(x)$ is now assumed to be convex.

Lemma 3.1 *For each k we have*

$$G_k(x) - G_k(x^k) \geq D_h(x, x^k) = g_{k+1}(x). \tag{3.10}$$

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

$$0 \in \partial f(x^k) + \nabla h(x^k) - \nabla h(x^{k-1}), \tag{3.11}$$

so that

$$\nabla h(x^{k-1}) = u^k + \nabla h(x^k), \tag{3.12}$$

for some u^k in $\partial f(x^k)$. 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 (3.12), to get

$$G_k(x) - G_k(x^k) = f(x) - f(x^k) - \langle u^k, x - x^k \rangle + D_h(x, x^k). \tag{3.13}$$

Therefore,

$$G_k(x) - G_k(x^k) \geq D_h(x, x^k),$$

since u^k is in $\partial f(x^k)$. ∎

3.2.4 Convergence of the PMA

From the discussion of the SUMMA we know that $\{f(x^k)\}$ is monotonically decreasing to $f(\hat{x})$. If $\mathcal{H} = \mathbb{R}^J$, if the sequence $\{x^k\}$ is bounded, and if $\hat{x}$ is unique, we can conclude that $\{x^k\} \to \hat{x}$.

For the remainder of this subsection we assume that $\mathcal{H} = \mathbb{R}^J$, in order

to make use of the results in Chpater 8. 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

$$G_k(\hat{x}) = f(\hat{x}) + D_h(\hat{x}, x^{k-1}). \tag{3.14}$$

From Equation (3.13) we have

$$G_k(\hat{x}) = G_k(x^k) + f(\hat{x}) - f(x^k) - \langle u^k, \hat{x} - x^k \rangle + D_h(\hat{x}, x^k). \tag{3.15}$$

Therefore,

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

$$f(x^k) - f(\hat{x}) + D_h(x^k, x^{k-1}) + f(\hat{x}) - f(x^k) - \langle u^k, \hat{x} - x^k \rangle. \tag{3.16}$$

It follows that the sequence $\{D_h(\hat{x}, x^k)\}$ is decreasing and that $\{f(x^k)\}$ converges to $f(\hat{x})$. 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 (see Chapter 8).

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

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

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

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

Now we summarize our results for the PMA. Let $f : \mathbb{R}^J \to (-\infty, +\infty]$ be closed, proper, and convex. 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 3.1 *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 3.2 *If $h(x)$ is a Bregman–Legendre function and $\hat{x}$ can be chosen in D, then $\{x^k\} \to x^*$, x^* in D, with $f(x^*) = f(\hat{x})$.*

3.2.5 The Non-Differentiable Case

In the discussion so far, we have assumed that the function $h(x)$ is differentiable; the gradient played a role in the definition of the Bregman distance $D_h(x, z)$. When $h(x)$ is not differentiable, a PMA is still available. In the non-differentiable case a Bregman distance is defined to be

$$D_h(x, z; p) = h(x) - h(z) - \langle p, x - z \rangle , \tag{3.17}$$

where p is a member of the subdifferential $\partial h(z)$. We begin the PMA by selecting initial vectors x^0 and $p^0 \in \partial h(x^0)$. Now the iterate x^k minimizes

$$G_k(x) = f(x) + D_h(x, x^{k-1}; p^{k-1}), \tag{3.18}$$

where p^{k-1} is a member of $\partial h(x^{k-1})$. Therefore,

$$0 \in \partial f(x^k) + \partial h(x^k) - p^{k-1}. \tag{3.19}$$

We assume that this equation can be solved and that there are $u^k \in \partial f(x^k)$ and $v^k \in \partial h(x^k)$ so that

$$v^k = p^{k-1} - u^k. \tag{3.20}$$

We then define $p^k = v^k$, so that

$$G_k(x) - G_k(x^k) =$$

$$D_f(x, x^k; u^k) + D_h(x, x^k; p^k) \geq D_h(x, x^k; p^k) = g_{k+1}(x). \tag{3.21}$$

Therefore, the SUMMA condition holds and the sequence $\{f(x^k)\}$ converges to $f(\hat{x})$.

3.3 The IPA

The IPA is a particular case of the PMA designed to overcome some of the computational obstacles encountered in the PMA [40, 47]. At the kth step of the PMA we must solve the equation

$$\nabla f(x^k) + \nabla h(x^k) = \nabla h(x^{k-1}) \tag{3.22}$$

for x^k, where, for notational convenience, we assume that both f and h are differentiable. Solving Equation (3.22) is probably not a simple matter,

however. In the IPA approach we begin not with $h(x)$, but with a convex differentiable function $a(x)$ such that $h(x) = a(x) - f(x)$ is convex. Equation (3.22) now reads

$$\nabla a(x^k) = \nabla a(x^{k-1}) - \nabla f(x^{k-1}), \tag{3.23}$$

and we choose $a(x)$ so that Equation (3.23) is easily solved. We turn now to several examples of the IPA.

3.4 Projected Gradient Descent

The problem now is to minimize $f : \mathbb{R}^J \to \mathbb{R}$, over the closed, nonempty convex set C, where f is convex and differentiable on $\mathbb{R}^J$. We assume now that the gradient operator ∇f is L-Lipschitz continuous; that is, for all x and y, we have

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|. \tag{3.24}$$

To employ the IPA approach, we let $0 < \gamma < \frac{1}{L}$ and select the function

$$a(x) = \frac{1}{2\gamma}\|x\|^2; \tag{3.25}$$

the upper bound on γ guarantees that the function $h(x) = a(x) - f(x)$ is convex. At the kth step we minimize

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

$$f(x) + \frac{1}{2\gamma}\|x - x^{k-1}\|^2 - D_f(x, x^{k-1}), \tag{3.26}$$

over $x \in C$. The solution x^k is in C and satisfies the inequality

$$\langle x^k - (x^{k-1} - \gamma\nabla f(x^{k-1})), c - x^k\rangle \geq 0, \tag{3.27}$$

for all $c \in C$. It follows then that

$$x^k = P_C(x^{k-1} - \gamma\nabla f(x^{k-1})); \tag{3.28}$$

here P_C denotes the orthogonal projection onto C. This is the projected gradient descent algorithm. For convergence we must require that f have certain additional properties needed for convergence of a PMA algorithm. Note that the auxiliary function

$$g_k(x) = \frac{1}{2\gamma}\|x - x^{k-1}\|^2 - D_f(x, x^{k-1}) \tag{3.29}$$

is unrelated to the set C, so is not used here to incorporate the constraint; it is used to provide a closed-form iterative scheme.

When $C = \mathbb{R}^J$ we have no constraint and the problem is simply to minimize f. Then the iterative algorithm becomes

$$x^k = x^{k-1} - \gamma \nabla f(x^{k-1}); \tag{3.30}$$

this is the gradient descent algorithm.

3.5 Relaxed Gradient Descent

In the gradient descent method we move away from the current x^{k-1} by the vector $\gamma \nabla f(x^{k-1})$. In relaxed gradient descent, the magnitude of the movement is reduced by a factor of α, where $\alpha \in (0, 1)$. Such relaxation methods are sometimes used to accelerate convergence. The relaxed gradient descent method can also be formulated as an AF method.

At the kth step we minimize

$$G_k(x) = f(x) + \frac{1}{2\gamma\alpha}\|x - x^{k-1}\|^2 - D_f(x, x^{k-1}), \tag{3.31}$$

obtaining

$$x^k = x^{k-1} - \alpha\gamma \nabla f(x^{k-1}). \tag{3.32}$$

3.6 Regularized Gradient Descent

In many applications the function to be minimized involves measured data, which is typically noisy, as well as some less than perfect model of how the measured data was obtained. In such cases, we may not want to minimize $f(x)$ exactly. In regularization methods we add to $f(x)$ another function that is designed to reduce sensitivity to noise and model error.

For example, suppose that we want to minimize

$$\alpha f(x) + \frac{1-\alpha}{2}\|x - p\|^2, \tag{3.33}$$

where p is chosen a priori. The regularized gradient descent algorithm for this problem can be put in the framework of an AF method.

At the kth step we minimize

$$G_k(x) = f(x) + \frac{1}{2\gamma\alpha}\|x - x^{k-1}\|^2 - \frac{1}{\alpha}(x, x^{k-1})] + \frac{1-\alpha}{2\gamma\alpha}\|x - p\|^2, \quad (3.34)$$

obtaining

$$x^k = \alpha(x^{k-1} - \gamma\nabla f(x^{k-1})) + (1-\alpha)p. \quad (3.35)$$

If we select $p = 0$ the iterative step becomes

$$x^k = \alpha(x^{k-1} - \gamma\nabla f(x^{k-1})). \quad (3.36)$$

3.7 The Projected Landweber Algorithm

The Landweber (LW) and projected Landweber (PLW) algorithms are special cases of projected gradient descent. The objective now is to minimize the function

$$f(x) = \frac{1}{2}\|Ax - b\|^2, \quad (3.37)$$

over $x \in \mathbb{R}^J$ or $x \in C$, where A is a real I by J matrix. The gradient of $f(x)$ is

$$\nabla f(x) = A^T(Ax - b) \quad (3.38)$$

and is L-Lipschitz continuous for $L = \rho(A^TA)$, the largest eiqenvalue of A^TA. The Bregman distance associated with $f(x)$ is

$$D_f(x, z) = \frac{1}{2}\|Ax - Az\|^2. \quad (3.39)$$

We let

$$a(x) = \frac{1}{2\gamma}\|x\|^2, \quad (3.40)$$

where $0 < \gamma < \frac{1}{L}$, so that the function $h(x) = a(x) - f(x)$ is convex.

At the kth step of the PLW we minimize

$$G_k(x) = f(x) + D_h(x, x^{k-1}) \quad (3.41)$$

over $x \in C$ to get

$$x^k = P_C(x^{k-1} - \gamma A^T(Ax^{k-1} - b)); \quad (3.42)$$

in the case of $C = \mathbb{R}^J$ we get the Landweber algorithm.

Chapter 4

Fixed-Point Methods

4.1 Introduction

We denote by $\mathcal{H}$ a real Hilbert space with inner product $\langle \cdot, \cdot \rangle$ and norm $\| \cdot \|$. We say that an operator $T : \mathcal{H} \to \mathcal{H}$ is *convergent* if, for every starting vector x^0, the sequence $\{x^k\}$ defined by $x^k = Tx^{k-1}$ converges weakly to a fixed point of T, whenever T has a fixed point. Fixed-point iterative methods are used to solve a variety of problems by selecting a convergent T for which the fixed points of T are solutions of the original problem. It is important, therefore, to identify properties of an operator T that guarantee that T is convergent.

An operator $T : \mathcal{H} \to \mathcal{H}$ is nonexpansive if, for all x and y in $\mathcal{H}$,

$$\|Tx - Ty\| \leq \|x - y\|. \tag{4.1}$$

Just being nonexpansive does not make T convergent, as the example $T = -Id$ shows; here Id is the identity operator. It doesn't take much, however, to convert a nonexpansive operator N into a convergent operator. Let $0 < \alpha < 1$ and $T = (1 - \alpha)Id + \alpha N$; then T is convergent. Such operators

are called *averaged* [5, 8, 43] and are convergent as a consequence of the Krasnosel'skii-Mann Theorem [15].

A operator $T : \mathcal{H} \to \mathcal{H}$ is *firmly nonexpansive* if, for all x and y in $\mathcal{H}$,

$$\langle Tx - Ty, x - y\rangle \geq \|Tx - Ty\|^2. \tag{4.2}$$

It is not hard to show that T is firmly nonexpansive if and only if $T = \frac{1}{2}(Id + N)$, for some nonexpansive operator N. Clearly, then, if T is firmly nonexpansive, T is averaged, and therefore T is nonexpansive, and all firmly nonexpansive operators are convergent. Also, T is firmly nonexpansive if and only if $G = Id - T$ is firmly nonexpansive. The Baillon–Haddad Theorem is the following.

Theorem 4.1 (The Baillon–Haddad Theorem) ([4], Corollaire 10]) *Let $f : \mathcal{H} \to \mathbb{R}$ be convex and Gâteaux differentiable on $\mathcal{H}$, and its gradient operator $T = \nabla f$ nonexpansive. Then f is Fréchet differentiable and T is firmly nonexpansive.*

In [4] this theorem appears as a corollary of a more general theorem concerning n-cyclically monotone operators in normed vector space. In [14] Bauschke and Combettes generalize the Baillon–Haddad Theorem, giving several additional conditions equivalent to the two in Theorem 4.1. Their proofs are not elementary.

The Baillon–Haddad Theorem provides an important link between convex optimization and fixed-point iteration. If $g : \mathcal{H} \to \mathbb{R}$ is a Gâteaux differentiable convex function and its gradient is L-Lipschitz continuous, that is,

$$\|\nabla g(x) - \nabla g(y)\| \leq L\|x - y\|, \tag{4.3}$$

for all x and y, then g is Fréchet differentiable and the gradient operator of the function $f = \frac{1}{L}g$ is nonexpansive. By the Baillon–Haddad Theorem the gradient operator of f is firmly nonexpansive. It follows that, for any $0 < \gamma < \frac{2}{L}$, the operator $Id - \gamma\nabla g$ is averaged, and therefore convergent. The class of averaged operators is closed to finite products, and P_C, the orthogonal projection onto a closed convex set C, is firmly nonexpansive. Therefore, the projected gradient-descent algorithm with the iterative step

$$x^{k+1} = P_C(x^k - \gamma\nabla g(x^k)) \tag{4.4}$$

converges weakly to a minimizer, over C, of the function g, whenever such minimizers exist.

4.2 The Krasnosel'skii-Mann-Opial Theorem

For any operator $T : \mathcal{H} \to \mathcal{H}$ that is averaged, weak convergence of the sequence $\{T^k x^0\}$ to a fixed point of T, whenever fixed points of T exist, is guaranteed by the Krasnosel'skii-Mann-Opial (KMO) Theorem [116, 128, 137]. The proof we present here is for the case of $\mathcal{H} = \mathbb{R}^J$; the proof is a bit more complicated for the infinite-dimensional case (see Theorem 5.14 in [14]).

Theorem 4.2 *Let $T : \mathbb{R}^J \to \mathbb{R}^J$ be α-averaged, for some $\alpha \in (0, 1)$. Then, for any x^0, the sequence $\{T^k x^0\}$ converges to a fixed point of T, whenever* Fix(T) *is nonempty.*

Proof: Let z be a fixed point of T. The identity in Equation (1.15) is the key to proving Theorem 4.2.

Using $Tz = z$ and $(I - T)z = 0$ and setting $G = I - T$ we have

$$||z - x^k||^2 - ||Tz - x^{k+1}||^2 = 2\langle Gz - Gx^k, z - x^k\rangle - ||Gz - Gx^k||^2. \tag{4.5}$$

Since G is $\frac{1}{2\alpha}$-ism, we have

$$||z - x^k||^2 - ||z - x^{k+1}||^2 \geq (\frac{1}{\alpha} - 1)||x^k - x^{k+1}||^2. \tag{4.6}$$

Consequently, the sequence $\{||z - x^k||\}$ is decreasing, the sequence $\{x^k\}$ is bounded, and the sequence $\{||x^k - x^{k+1}||\}$ converges to zero. Let x^* be a cluster point of $\{x^k\}$. Then we have $Tx^* = x^*$, so we may use x^* in place of the arbitrary fixed point z. It follows then that the sequence $\{||x^* - x^k||\}$ is decreasing. Since a subsequence converges to zero, the entire sequence converges to zero. ∎

A version of the KMO Theorem 4.2, with variable coefficients, appears in Reich's paper [139].

An operator T is said to be *asymptotically regular* if, for any x, the sequence $\{\|T^k x - T^{k+1}x\|\}$ converges to zero. The proof of the KMO Theorem 4.2 involves showing that any averaged operator with fixed points is asymptotically regular. In [137] Opial generalizes the KMO Theorem, proving that, if T is nonexpansive and asymptotically regular, then the sequence $\{T^k x\}$ converges to a fixed point of T, whenever fixed points exist, for any x.

Note that, in the KMO Theorem, we assumed that T is α-averaged, so that $G = I - T$ is ν-ism, for some $\nu > \frac{1}{2}$. But we actually used a somewhat weaker condition on G; we required only that

$$\langle Gz - Gx, z - x\rangle \geq \nu\|Gz - Gx\|^2$$

for z such that $Gz = 0$. This weaker property is called *weakly* ν-ism.

We showed previously that the projected gradient descent (PGD) algorithm is a particular case of the PMA, for $0 < \gamma < \frac{1}{L}$. Convergence of the PGD algorithm then follows, for those functions f that satisfy the conditions needed for convergence of the PMA. Now that we have Theorem 4.2 we can say more. We know now that the operator $T = P_C(I - \gamma\nabla f)$ is averaged, and therefore convergent, for $0 < \gamma < \frac{2}{L}$. We now know that the CQ algorithm converges whenever there are fixed points, as do all particular cases of the CQ algorithm, such as the Landweber and projected Landweber algorithms.

The CQ algorithm is a particular case of the more general *forward-backward splitting* (FBS) algorithm. Before we can present the FBS algorithm we need to discuss the Moreau envelope and the Fenchel conjugate.

4.3 The Fenchel Conjugate

The duality between convex functions on $\mathcal{H}$ and their tangent hyperplanes is made explicit through the Legendre-Fenchel transformation. Initially, we take $f : \mathcal{H} \to [-\infty, +\infty]$, without any additional assumptions on f.

4.3.1 The Fenchel Conjugate

We say that a function $h(x) : \mathcal{H} \to \mathbb{R}$ is *affine* if it has the form $h(x) = \langle a, x\rangle - \gamma$, for some vector a and scalar γ. If $\gamma = 0$, then we call the function *linear*. A function such as $f(x) = 5x + 2$ is commonly called a linear function in algebra classes, but, according to our definition, it should be called an affine function.

For each fixed vector a in $\mathcal{H}$, the affine function $h(x) = \langle a, x\rangle - \gamma$ is beneath the function $f(x)$ if $f(x) - h(x) \geq 0$, for all x; that is,

$$f(x) - \langle a, x\rangle + \gamma \geq 0,$$

or

$$\gamma \geq \langle a, x\rangle - f(x). \tag{4.7}$$

This leads us to the following definition, involving the supremum of the right side of the inequality in (4.7), for each fixed a.

Definition 4.1 *The* conjugate function *associated with f is the function*

$$f^*(a) = \sup_{x\in\mathcal{H}}(\langle a, x\rangle - f(x)). \tag{4.8}$$

For each fixed a, the value $f^*(a)$ is the smallest value of γ for which the affine function $h(x) = \langle a, x\rangle - \gamma$ is beneath $f(x)$ for all $x \in \mathcal{H}$. The passage from f to f^* is the *Legendre–Fenchel Transformation*. If f^* is proper, then so is f. The function f^* is always convex, since

$$f^*(a) = \sup_{(x,\eta)\in \mathrm{epi}(f)} \{\langle a, x\rangle - \eta\},$$

and the supremum of a family of affine functions is convex. The *epigraph* of a function $f : \mathcal{H} \to [-\infty, +\infty]$, denoted $\mathrm{epi}(f)$, is the set

$$\mathrm{epi}(f) = \{(x, \eta) | f(x) \leq \eta\}.$$

For example, suppose that $f(x) = \frac{1}{2}x^2$. The function $h(x) = ax + b$ is beneath $f(x)$ for all x if

$$ax + b \leq \frac{1}{2}x^2,$$

for all x. Equivalently,

$$b \leq \frac{1}{2}x^2 - ax,$$

for all x. Then b must not exceed the minimum of the right side, which is $-\frac{1}{2}a^2$ and occurs when $x - a = 0$, or $x = a$. Therefore, we have

$$\gamma = -b \geq \frac{1}{2}a^2.$$

The smallest value of γ for which this is true is $\gamma = \frac{1}{2}a^2$, so we have $f^*(a) = \frac{1}{2}a^2$.

4.3.2 The Conjugate of the Conjugate

Now we repeat this process with $f^*(a)$ in the role of $f(x)$. For each fixed vector x, the affine function $c(a) = \langle a, x\rangle - \gamma$ is beneath the function $f^*(a)$ if $f^*(a) - c(a) \geq 0$, for all $a \in \mathcal{H}$; that is,

$$f^*(a) - \langle a, x\rangle + \gamma \geq 0,$$

or

$$\gamma \geq \langle a, x\rangle - f^*(a). \tag{4.9}$$

This leads us to the following definition, involving the supremum of the right side of the inequality in (4.9), for each fixed x.

Definition 4.2 *The conjugate function associated with f^* is the function*

$$f^{**}(x) = \sup_a(\langle a, x\rangle - f^*(a)). \tag{4.10}$$

For each fixed x, the value $f^{**}(x)$ is the smallest value of γ for which the affine function $c(a) = \langle a, x\rangle - \gamma$ is beneath $f^*(a)$ for all a.

If f is closed and convex, we have $(f^*)^* = f^{**} = f$. Applying the Separation Theorem to the epigraph of the closed, proper, convex function $f(x)$, it can be shown ([143], Theorem 12.1) that $f(x)$ is the point-wise supremum of all the affine functions beneath $f(x)$; that is,

$$f(x) = \sup_{a,\gamma}\{h(x)|f(x) \geq h(x)\}.$$

Therefore,

$$f(x) = \sup_{a}\Big(\langle a, x\rangle - f^*(a)\Big).$$

This says that

$$f^{**}(x) = f(x). \tag{4.11}$$

If $f(x)$ is a differentiable function, then, for each fixed a, the function

$$g(x) = \langle a, x\rangle - f(x)$$

will attain its minimum if and only if

$$0 = \nabla g(x) = a - \nabla f(x),$$

which says that $a = \nabla f(x)$.

4.3.3 Some Examples of Conjugate Functions

- The exponential function $f(x) = \exp(x) = e^x$ has conjugate

$$\exp^*(a) = a\log a - a, \tag{4.12}$$

  if $a > 0$, 0 if $a = 0$, and $+\infty$ if $a < 0$.

- The function $f(x) = -\log x$, for $x > 0$, has the conjugate function $f^*(a) = -1 - \log(-a)$, for $a < 0$.

- The function $f(x) = \frac{|x|^p}{p}$ has conjugate $f^*(a) = \frac{|a|^q}{q}$, where $p > 0$, $q > 0$, and $\frac{1}{p} + \frac{1}{q} = 1$. Therefore, the function $f(x) = \frac{1}{2}\|x\|^2$ is its own conjugate, that is, $f^*(a) = \frac{1}{2}\|a\|^2$.

- Let A be a real symmetric positive-definite matrix and

$$f(x) = \frac{1}{2}\langle Ax, x\rangle.$$

  Then

$$f^*(a) = \frac{1}{2}\langle A^{-1}a, a\rangle.$$

- Let $i_C(x)$ be the *indicator function* of the closed convex set C, that is, $i_C(x) = 0$, if $x \in C$, and ∞ otherwise.Then

$$i_C^*(a) = \sup_{x \in C} \langle a, x \rangle,$$

which is the *support function* of the set C, usually denoted $\sigma_C(a)$.

- Let $C \subseteq \mathbb{R}^J$ be nonempty, closed and convex. The *gauge function* of C is

$$\gamma_C(x) = \inf\{\lambda \geq 0 \,|x \in \lambda C\}.$$

If $C = B$, the unit ball of $\mathbb{R}^J$, then $\gamma_B(x) = \|x\|$. For each C define the *polar set* for C by

$$C^0 = \{z | \langle z, c \rangle \leq 1, \text{ for all } c \in C\}.$$

Then

$$\gamma_C^* = \iota_{C^0}.$$

- Let $C = \{x| \ ||x|| \leq 1\}$, so that the function $\phi(a) = \|a\|_2$ satisfies

$$\phi(a) = \sup_{x \in C} \langle a, x \rangle.$$

Then

$$\phi(a) = \sigma_C(a) = i_C^*(a).$$

Therefore,

$$\phi^*(x) = \sigma_C^*(x) = i_C^{**}(x) = i_C(x).$$

4.3.4 Conjugates and Subgradients

We know from the definition of $f^*(a)$ that

$$f^*(a) \geq \langle a, z \rangle \ - f(z),$$

for all z, and, moreover, $f^*(a)$ is the supremum of these values, taken over all z. If a is a member of the subdifferential $\partial f(x)$, then, for all z, we have

$$f(z) \geq f(x) + \langle a, z - x \rangle,$$

so that

$$\langle a, x \rangle - f(x) \geq \langle a, z \rangle - f(z).$$

It follows that

$$f^*(a) = \langle a, x \rangle \ - f(x),$$

so that

$$f(x) + f^*(a) = \langle a, x \rangle.$$

If $f(x)$ is a differentiable convex function, then a is in the subdifferential $\partial f(x)$ if and only if $a = \nabla f(x)$. Then we can say

$$f(x) + f^*(\nabla f(x)) = \langle \nabla f(x), x \rangle. \tag{4.13}$$

If $a = \nabla f(x_1)$ and $a = \nabla f(x_2)$, then the function

$$g(x) = \langle a, x \rangle - f(x)$$

attains its maximum value at $x = x_1$ and at $x = x_2$, so that

$$f^*(a) = \langle a, x_1 \rangle - f(x_1) = \langle a, x_2 \rangle - f(x_2).$$

Let us denote by $x = (\nabla f)^{-1}(a)$ any x for which $\nabla f(x) = a$. Then the conjugate of the differentiable function $f : \mathcal{H} \to \mathbb{R}$ can then be defined as follows [143]. Let D be the image of $\mathcal{H}$ under the mapping ∇f. Then, for all $a \in D$, define

$$f^*(a) = \langle a, (\nabla f)^{-1}(a) \rangle - f((\nabla f)^{-1}(a)). \tag{4.14}$$

The formula in Equation (4.14) is also called the Legendre Transform.

4.4 The Forward-Backward Splitting Algorithm

The *forward-backward splitting* (FBS) methods [78, 51] form a broad class of SUMMA algorithms closely related the IPA. Note that minimizing $G_k(x)$ in Equation (3.3) over $x \in C$ is equivalent to minimizing

$$G_k(x) = \iota_C(x) + f(x) + D_h(x, x^{k-1}) \tag{4.15}$$

over all $x \in \mathbb{R}^J$, where $\iota_C(x) = 0$ for $x \in C$ and $\iota_C(x) = +\infty$ otherwise. This suggests a more general iterative algorithm, the FBS.

Suppose that we want to minimize the function $f_1(x) + f_2(x)$, where both functions are convex and $f_2(x)$ is differentiable with its gradient L-Lipschitz continuous in the Euclidean norm, by which we mean that

$$\|\nabla f_2(x) - \nabla f_2(y)\| \leq L\|x - y\|, \tag{4.16}$$

for all x and y. At the kth step of the FBS algorithm we obtain x^k by minimizing

$$G_k(x) = f_1(x) + f_2(x) + \frac{1}{2\gamma}\|x - x^{k-1}\|^2 - D_{f_2}(x, x^{k-1}), \tag{4.17}$$

over all $x \in \mathbb{R}^J$, where $0 < \gamma < \frac{1}{L}$.

4.5 Moreau's Proximity Operators

Following Combettes and Wajs [78], we say that the *Moreau envelope* of index $\gamma > 0$ of the closed, proper, convex function $f : \mathcal{H} \to (-\infty, \infty]$, or the Moreau envelope of the function γf, is the continuous, convex function

$$^{\gamma}f(x) = \text{env}_{\gamma f}(x) = \inf_{y \in \mathcal{H}} \{f(y) + \frac{1}{2\gamma}||x - y||^2\}; \tag{4.18}$$

see also Moreau [131, 132, 133]. In Chapter 7 we shall denote $\text{env}_f(x)$ by $m_f(x)$.

In Rockafellar's book [143] and elsewhere, it is shown that the infimum is attained at a unique y, usually denoted $\text{prox}_{\gamma f}(x)$. As we shall see in Proposition 7.2, the proximity operators $\text{prox}_{\gamma f}(\cdot)$ are firmly nonexpansive; indeed, the proximity operator prox_f is the resolvent of the maximal monotone operator $B(x) = \partial f(x)$ and all such resolvent operators are firmly nonexpansive [29]. Proximity operators also generalize the orthogonal projections onto closed, convex sets. Consider the function $f(x) = \iota_C(x)$, the *indicator function* of the closed, convex set C, taking the value zero for x in C, and $+\infty$ otherwise. Then $\text{prox}_{\gamma f}(x) = P_C(x)$, the orthogonal projection of x onto C. The following characterization of $x = \text{prox}_f(z)$ is quite useful: $x = \text{prox}_f(z)$ if and only if $z - x \in \partial f(x)$ (see Proposition 7.1).

4.6 The FBS Algorithm

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 [78].

Let $f : \mathbb{R}^J \to \mathbb{R}$ be convex, with $f = f_1 + f_2$, both convex, f_2 differentiable, and ∇f_2 L-Lipschitz continuous. The iterative step of the FBS algorithm is

$$x^k = \text{prox}_{\gamma f_1}\Big(x^{k-1} - \gamma \nabla f_2(x^{k-1})\Big). \tag{4.19}$$

As we shall show, convergence of the sequence $\{x^k\}$ to a solution can be established, if γ is chosen to lie within the interval $(0, 1/L]$.

4.7 Convergence of the FBS algorithm

We shall prove convergence of the FBS algorithm in two ways. First, we do so using the PMA framework. After that, we prove a somewhat stronger convergence result using Theorem 4.2.

Let $f : \mathbb{R}^J \to \mathbb{R}$ be convex, with $f = f_1 + f_2$, both convex, f_2 differentiable, and ∇f_2 L-Lipschitz continuous. Let $\{x^k\}$ be defined by Equation (4.19) and let $0 < \gamma \leq 1/L$.

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

$$G_k(x) = f(x) + \frac{1}{2\gamma}\|x - x^{k-1}\|^2 - D_{f_2}(x, x^{k-1}), \tag{4.20}$$

where

$$D_{f_2}(x, x^{k-1}) = f_2(x) - f_2(x^{k-1}) - \langle \nabla f_2(x^{k-1}), x - x^{k-1} \rangle. \tag{4.21}$$

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.

The auxiliary function

$$g_k(x) = \frac{1}{2\gamma}\|x - x^{k-1}\|^2 - D_{f_2}(x, x^{k-1}) \tag{4.22}$$

can be rewritten as

$$g_k(x) = D_h(x, x^{k-1}), \tag{4.23}$$

where

$$h(x) = \frac{1}{2\gamma}\|x\|^2 - f_2(x). \tag{4.24}$$

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

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

$$\langle \nabla h(x) - \nabla h(y), x - y \rangle \geq 0, \tag{4.25}$$

for all x and y. This is equivalent to

$$\frac{1}{\gamma}\|x - y\|^2 - \langle \nabla f_2(x) - \nabla f_2(y), x - y \rangle \geq 0. \tag{4.26}$$

Since ∇f_2 is L-Lipschitz, the inequality (4.26) holds for $0 < \gamma \leq 1/L$.

Lemma 4.1 *The x^k that minimizes $G_k(x)$ over x is given by Equation (4.19).*

Proof: We know that x^k minimizes $G_k(x)$ if and only if

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

or, equivalently,

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

Consequently,

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

∎

Theorem 4.3 *The sequence $\{x^k\}$ converges to a minimizer of the function $f(x)$, whenever minimizers exist.*

Proof: A relatively simple calculation shows that

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

$$\Big(f_1(x) - f_1(x^k) - \frac{1}{\gamma}\langle (x^{k-1} - \gamma \nabla f_2(x^{k-1})) - x^k, x - x^k\rangle\Big). \tag{4.27}$$

Since

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

it follows that

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

Therefore,

$$G_k(x) - G_k(x^k) \geq \frac{1}{2\gamma}\|x - x^k\|^2 \geq g_{k+1}(x). \tag{4.28}$$

Therefore, the SUMMA inequality 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

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

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

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

it follows that the sequence $\{x^k\}$ is bounded. Therefore, we may select a subsequence $\{x^{k_n}\}$ converging to some x^{**}, with $\{x^{k_n-1}\}$ converging to some x^*, and therefore $f(x^*) = f(x^{**}) = f(\hat{x})$.

Replacing the generic $\hat{x}$ with x^{**}, we find that $\{G_k(x^{**}) - G_k(x^k)\}$ is decreasing to zero. From the inequality in (4.28), we conclude that the sequence $\{\|x^* - x^k\|^2\}$ converges to zero, and so $\{x^k\}$ converges to x^*. This completes the proof of the theorem. ∎

Now we prove convergence of the FBS algorithm using Theorem 4.2. By Proposition 7.2 the operator prox_f is fne. Therefore, the operator $T = \text{prox}_{\gamma f}(I - \gamma\nabla f_2)$ is averaged, for $0 < \gamma < \frac{2}{L}$. According to Theorem 4.2, the FBS algorithm converges to a fixed point whenever fixed points exist.

4.8 Some Examples

We present some examples to illustrate the application of Theorem 4.2.

4.8.1 Projected Gradient Descent

Let C be a nonempty, closed convex subset of $\mathbb{R}^J$ 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 (4.19) becomes

$$x^k = P_C\Big(x^{k-1} - \gamma\nabla f_2(x^{k-1})\Big). \tag{4.29}$$

The sequence $\{x^k\}$ converges to a minimizer of f_2 over $x \in C$, whenever such minimizers exist, for $0 < \gamma \leq 1/L$.

4.8.2 The CQ Algorithm

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

$$f_2(x) = \frac{1}{2}\|P_Q Ax - Ax\|^2 \tag{4.30}$$

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

$$\nabla f_2(x) = A^T(I - P_Q)Ax. \tag{4.31}$$

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

$$x^k = P_C\Big(x^{k-1} - \gamma A^T(I - P_Q)Ax^{k-1}\Big); \tag{4.32}$$

this iterative method was called the CQ-algorithm in [42, 43]. The sequence $\{x^k\}$ converges to a solution whenever f_2 has a minimum on the set C, for $0 < \gamma \leq 1/L$.

Chapter 5

The SMART and EMML Algorithms

We turn now to iterative algorithms involving nonnegative vectors and matrices. For such algorithms the two-norm will not play a major role. Instead, the Kullback-Leibler, or cross-entropy, distance will be our primary tool. Our main examples are the simultaneous multiplicative algebraic reconstruction technique (SMART), the expectation maximization maximum likelihood (EMML) algorithms, and various related methods.

5.1 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

$$x_j^k = x_j^{k-1} \exp\left(s_j^{-1} \sum_{i=1}^{I} P_{ij} \log \frac{y_i}{(Px^{k-1})_i} \right). \tag{5.1}$$

5.2 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

$$x_j^k = x_j^{k-1} s_j^{-1} \Big(\sum_{i=1}^{I} P_{ij} \frac{y_i}{(Px^{k-1})_i} \Big). \tag{5.2}$$

5.3 The EMML and the SMART as AM

In [32] the SMART was derived using the following alternating minimization (AM) approach.

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

$$r(x)_{ij} = x_j P_{ij} y_i / (Px)_i, \tag{5.3}$$

and

$$q(x)_{ij} = x_j P_{ij}. \tag{5.4}$$

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 $KL(y, Px) = KL(r(x), q(x))$. It follows from the identities to be discussed in the next section that the SMART can also be formulated as a particular case of SUMMA.

5.4 The SMART as SUMMA

We show now that the SMART is a particular case of SUMMA; Lemma 2.1 is helpful in that regard. For notational convenience, we assume, for the remainder of this chapter, that $s_j = 1$ for all j; if this is not the case initially, we can rescale both P and x without changing Px. From the identities established for the SMART in [32] and reviewed later in this chapter, we know that the iterative step of SMART can be expressed as follows: minimize the function

$$G_k(x) = KL(Px, y) + KL(x, x^{k-1}) - KL(Px, Px^{k-1}) \tag{5.5}$$

to get x^k. According to Lemma 2.1, 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 $\mathbb{R}^J$. Each x^k is a positive vector.

It was shown in [32] that

$$G_k(x) = G_k(x^k) + KL(x, x^k), \tag{5.6}$$

from which it follows immediately that the SMART is in the SUMMA class.

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 [32] 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.

5.5 The SMART as PMA

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

$$F(x) = \sum_{j=1}^{J} x_j \log x_j. \tag{5.7}$$

Then

$$D_F(x,z) = KL(x,z). \tag{5.8}$$

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

$$D_f(x,z) = KL(Px,Pz). \tag{5.9}$$

Lemma 5.1 $D_F(x,z) \geq D_f(x,z)$.

Proof: We have

$$\begin{aligned} D_F(x,z) &\geq \sum_{j=1}^{J} KL(x_j,z_j) \geq \sum_{j=1}^{J}\sum_{i=1}^{I} KL(P_{ij}x_j,P_{ij}z_j) \\ &\geq \sum_{i=1}^{I} KL((Px)_i,(Pz)_i) = KL(Px,Pz). \end{aligned} \tag{5.10}$$

∎

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

$$f(x) + D_h(x,x^{k-1}). \tag{5.11}$$

So the SMART is a particular case of the PMA.

The function $h(x) = F(x) - f(x)$ is finite on $D = \mathbb{R}^J_+$, the nonnegative orthant of $\mathbb{R}^J$, and differentiable on its 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 (3.16) 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 5.1 *In the consistent case the SMART converges to the unique nonnegative solution of $y = Px$ for which the distance $\sum_{j=1}^{J} KL(x_j, x_j^0)$ is minimized. In the inconsistent case it converges to the unique nonnegative minimizer of the distance $KL(Px, y)$ for which $\sum_{j=1}^{J} KL(x_j, x_j^0)$ 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.*

5.6 Using KL Projections

For each $i = 1, 2, ..., I$, let H_i be the hyperplane

$$H_i = \{z | (Pz)_i = y_i\}. \tag{5.12}$$

The KL projection of a given positive x onto H_i is the z in H_i that minimizes the KL distance $KL(z, x)$. Generally, the KL projection onto H_i cannot be expressed in closed form. However, the z in H_i that minimizes the weighted KL distance

$$\sum_{j=1}^{J} P_{ij} KL(z_j, x_j) \tag{5.13}$$

is $T_i(x)$ given by

$$T_i(x)_j = x_j y_i / (Px)_i. \tag{5.14}$$

Both the SMART and the EMML can be described in terms of the T_i.

The iterative step of the SMART algorithm can be expressed as

$$x_j^k = \prod_{i=1}^{I} (T_i(x^{k-1})_j)^{P_{ij}}. \tag{5.15}$$

We see that x_j^k is a weighted geometric mean of the terms $T_i(x^{k-1})_j$.

The iterative step of the EMML algorithm can be expressed as

$$x_j^k = \sum_{i=1}^{I} P_{ij} T_i(x^{k-1})_j. \tag{5.16}$$

We see that x_j^k is a weighted arithmetic mean of the terms $T_i(x^{k-1})_j$, using the same weights as in the case of SMART.

5.7 The MART and EMART Algorithms

The MART algorithm has the iterative step

$$x_j^k = x_j^{k-1} \left(\frac{y_i}{(Px^{k-1})_i} \right)^{P_{ij} m_i^{-1}}, \tag{5.17}$$

where $i = (k-1)(\mathrm{mod}\, I) + 1$ and

$$m_i = \max\{P_{ij} | j = 1, 2, ..., J\}. \tag{5.18}$$

When there are nonnegative solutions of the system $y = Px$, the sequence $\{x^k\}$ converges to the solution x that minimizes $KL(x, x^0)$ [35, 36, 37]. We can express the MART in terms of the weighted KL projections $T_i(x^{k-1})$;

$$x_j^k = (x_j^{k-1})^{1 - P_{ij} m_i^{-1}} (T_i(x^{k-1})_j)^{P_{ij} m_i^{-1}}. \tag{5.19}$$

We see then that the iterative step of the MART is a relaxed weighted KL projection onto H_i, and a weighted geometric mean of the current x_j^{k-1} and $T_i(x^{k-1})_j$. The expression for the MART in Equation (5.19) suggests a somewhat simpler iterative algorithm involving a weighted arithmetic mean of the current x_j^{k-1} and $T_i(x^{k-1})_j$; this is the EMART algorithm.

The iterative step of the EMART algorithm is

$$x_j^k = (1 - P_{ij} m_i^{-1}) x_j^{k-1} + P_{ij} m_i^{-1} T_i(x^{k-1})_j. \tag{5.20}$$

Whenever the system $y = Px$ has nonnegative solutions, the EMART sequence $\{x^k\}$ converges to a nonnegative solution, but nothing further is known about this solution. One advantage that the EMART has over the MART is the substitution of multiplication for exponentiation.

Block-iterative versions of SMART and EMML have also been investigated; see [35, 36, 37] and the references therein.

5.8 Extensions of MART and EMART

As we have seen, the iterative steps of the MART and the EMART are relaxed weighted KL projections onto the hyperplane H_i, resulting in vectors that are not within H_i. This suggests variants of MART and EMART in which, at the end of each iterative step, a further weighted KL projection onto H_i is performed. In other words, for MART and EMART the new vector would be $T_i(x^k)$, instead of x^k as given by Equations (5.17) and (5.20), respectively. Research into the properties of these new algorithms is ongoing.

5.9 Convergence of the SMART and EMML

In this section we prove convergence of the SMART and EMML algorithms through a series of exercises. For both algorithms we begin with an arbitrary positive vector x^0. The iterative step for the EMML method is

$$x_j^k = (x^{k-1})'_j = x_j^{k-1} \sum_{i=1}^{I} P_{ij} \frac{y_i}{(Px^{k-1})_i}. \tag{5.21}$$

The iterative step for the SMART is

$$x_j^m = (x^{m-1})''_j = x_j^{m-1} \exp\Big(\sum_{i=1}^{I} P_{ij} \log \frac{y_i}{(Px^{m-1})_i} \Big). \tag{5.22}$$

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

5.9.1 Pythagorean Identities for the KL Distance

The SMART and EMML iterative algorithms are best derived using the principle of *alternating minimization*, according to which the distances $KL(r(x), q(z))$ and $KL(q(x), r(z))$ are minimized, first with respect to the variable x and then with respect to the variable z. Although the KL distance is not Euclidean, and, in particular, not even symmetric, there are analogues of Pythagoras' theorem that play important roles in the convergence proofs.

Ex. 5.1 *Establish the following Pythagorean identities:*

$$KL(r(x), q(z)) = KL(r(z), q(z)) + KL(r(x), r(z)); \tag{5.23}$$

$$KL(r(x), q(z)) = KL(r(x), q(x')) + KL(x', z), \tag{5.24}$$

for

$$x'_j = x_j \sum_{i=1}^{I} P_{ij} \frac{y_i}{(Px)_i}; \tag{5.25}$$

$$KL(q(x), r(z)) = KL(q(x), r(x)) + KL(x, z) - KL(Px, Pz); \tag{5.26}$$

$$KL(q(x), r(z)) = KL(q(z''), r(z)) + KL(x, z''), \tag{5.27}$$

for

$$z''_j = z_j \exp\left(\sum_{i=1}^{I} P_{ij} \log \frac{y_i}{(Pz)_i}\right). \tag{5.28}$$

Note that it follows from Equation (2.7) that $KL(x, z) - KL(Px, Pz) \geq 0$.

5.9.2 Convergence Proofs

We shall prove convergence of the SMART and EMML algorithms through a series of exercises.

Ex. 5.2 *Show that, for* $\{x^k\}$ *given by Equation (5.21),* $\{KL(y, Px^k)\}$ *is decreasing and* $\{KL(x^{k+1}, x^k)\} \to 0$. *Show that, for* $\{x^m\}$ *given by Equation (5.22),* $\{KL(Px^m, y)\}$ *is decreasing and* $\{KL(x^m, x^{m+1})\} \to 0$. *Hint: Use* $KL(r(x), q(x)) = KL(y, Px)$, $KL(q(x), r(x)) = KL(Px, y)$, *and the Pythagorean identities.*

Ex. 5.3 *Show that the EMML sequence* $\{x^k\}$ *is bounded by showing*

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

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

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

Ex. 5.4 *Show that* $(x^*)' = x^*$ *for any cluster point* x^* *of the EMML sequence* $\{x^k\}$ *and that* $(x^*)'' = x^*$ *for any cluster point* x^* *of the SMART sequence* $\{x^m\}$. *Hint: Use* $\{KL(x^{k+1}, x^k)\} \to 0$ *and* $\{KL(x^m, x^{m+1})\} \to 0$.

Ex. 5.5 *Let* $\hat{x}$ *and* $\tilde{x}$ *minimize* $KL(y, Px)$ *and* $KL(Px, y)$, *respectively, over all* $x \geq 0$. *Then,* $(\hat{x})' = \hat{x}$ *and* $(\tilde{x})'' = \tilde{x}$. *Hint: Apply Pythagorean identities to* $KL(r(\hat{x}), q(\hat{x}))$ *and* $KL(q(\tilde{x}), r(\tilde{x}))$.

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

Ex. 5.6 *For the EMML sequence $\{x^k\}$ with cluster point x^* and $\hat{x}$ as defined previously, we have the double inequality*

$$KL(\hat{x}, x^k) \geq KL(r(\hat{x}), r(x^k)) \geq KL(\hat{x}, x^{k+1}), \tag{5.29}$$

from which we conclude that the sequence $\{KL(\hat{x}, x^k)\}$ is decreasing and $KL(\hat{x}, x^*) < +\infty$. Hint: For the first inequality calculate $KL(r(\hat{x}), q(x^k))$ in two ways. For the second one, use $(x)'_j = \sum_{i=1}^{I} r(x)_{ij}$ and Lemma 2.1.

Ex. 5.7 *Show that, for the SMART sequence $\{x^m\}$ with cluster point x^* and $\tilde{x}$ as defined previously, we have*

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

$$KL(P\tilde{x}, Px^m) + KL(x^{m+1}, x^m) - KL(Px^{m+1}, Px^m), \tag{5.30}$$

and so $KL(P\tilde{x}, Px^*) = 0$, the sequence $\{KL(\tilde{x}, x^m)\}$ is decreasing and $KL(\tilde{x}, x^*) < +\infty$. Hint: Expand $KL(q(\tilde{x}), r(x^m))$ using the Pythagorean identities.

Ex. 5.8 *For x^* a cluster point of the EMML sequence $\{x^k\}$ we have $KL(y, Px^*) = KL(y, P\hat{x})$. Therefore, x^* is a nonnegative minimizer of $KL(y, Px)$. Consequently, the sequence $\{KL(x^*, x^k)\}$ converges to zero, and so $\{x^k\} \to x^*$. Hint: Use the double inequality of Equation (5.29) and $KL(r(\hat{x}), q(x^*))$.*

Ex. 5.9 *For x^* a cluster point of the SMART sequence $\{x^m\}$ we have $KL(Px^*, y) = KL(P\tilde{x}, y)$. Therefore, x^* is a nonnegative minimizer of $KL(Px, y)$. Consequently, the sequence $\{KL(x^*, x^m)\}$ converges to zero, and so $\{x^m\} \to x^*$. Moreover,*

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

for all $\tilde{x}$ as before. Hints: Use Exercise 5.7. For the final assertion use the fact that the difference $KL(\tilde{x}, x^m) - KL(\tilde{x}, x^{m+1})$ is independent of the choice of $\tilde{x}$, since it depends only on $Px^* = P\tilde{x}$. Now sum over the index m.

5.10 Regularization

The "night sky" phenomenon that occurs in nonnegatively constrained least-squares also happens with methods based on the Kullback-Leibler distance, such as MART, EMML and SMART, requiring some sort of regularization.

5.10.1 The "Night-Sky" Problem

As we saw previously, the sequence $\{x^k\}$ generated by the EMML iterative step in Equation (5.2) converges to a nonnegative minimizer $\hat{x}$ of $f(x) = KL(y, Px)$, and we have

$$\hat{x}_j = \hat{x}_j \sum_{i=1}^{I} P_{ij} \frac{y_i}{(P\hat{x})_i}, \tag{5.31}$$

for all j. We consider what happens when there is no nonnegative solution of the system $y = Px$.

For those values of j for which $\hat{x}_j > 0$, we have

$$1 = \sum_{i=1}^{I} P_{ij} = \sum_{i=1}^{I} P_{ij} \frac{y_i}{(P\hat{x})_i}. \tag{5.32}$$

Now let Q be the I by K matrix obtained from P by deleting rows j for which $\hat{x}_j = 0$. If Q has full rank and $K \geq I$, then Q^T is one-to-one, so that $1 = \frac{y_i}{(P\hat{x})_i}$ for all i, or $y = P\hat{x}$. But we are assuming that there is no nonnegative solution of $y = Px$. Consequently, we must have $K < I$ and $I - K$ of the entries of $\hat{x}$ are zero.

5.11 Modifying the KL distance

The SMART, EMML and their block-iterative versions are based on the Kullback-Leibler distance between nonnegative vectors and require that the solution sought be a nonnegative vector. To impose more general constraints on the entries of x we derive algorithms based on shifted KL distances, also called Fermi-Dirac generalized entropies.

For a fixed real vector u, the shifted KL distance $KL(x - u, z - u)$ is defined for vectors x and z having $x_j \geq u_j$ and $z_j \geq u_j$. Similarly, the

shifted distance $KL(v-x, v-z)$ applies only to those vectors x and z for which $x_j \leq v_j$ and $z_j \leq v_j$. For $u_j \leq v_j$, the combined distance

$$KL(x-u, z-u) + KL(v-x, v-z)$$

is restricted to those x and z whose entries x_j and z_j lie in the interval $[u_j, v_j]$. Our objective is to mimic the derivation of the SMART, EMML and RBI methods, replacing KL distances with shifted KL distances, to obtain algorithms that enforce the constraints $u_j \leq x_j \leq v_j$, for each j. The algorithms that result are the ABMART and ABEMML block-iterative methods. These algorithms were originally presented in [38], in which the vectors u and v were called a and b, hence the names of the algorithms. As previously, we shall assume that the entries of the matrix P are nonnegative. We shall denote by B_n, $n = 1, ..., N$ a partition of the index set $\{i = 1, ..., I\}$ into blocks. For $k = 0, 1, ...$ let $n(k) = k(\mathrm{mod}\, N) + 1$.

The projected Landweber algorithm can also be used to impose the restrictions $u_j \leq x_j \leq v_j$; however, the projection step in that algorithm is implemented by clipping, or setting equal to u_j or v_j values of x_j that would otherwise fall outside the desired range. The result is that the values u_j and v_j can occur more frequently than may be desired. One advantage of the AB methods is that the values u_j and v_j represent barriers that can only be reached in the limit and are never taken on at any step of the iteration.

5.12 The ABMART Algorithm

We assume that $(Pu)_i \leq y_i \leq (Pv)_i$ and seek a solution of $Px = y$ with $u_j \leq x_j \leq v_j$, for each j. The algorithm begins with an initial vector x^0 satisfying $u_j \leq x_j^0 \leq v_j$, for each j. Having calculated x^k, we take

$$x_j^{k+1} = \alpha_j^k v_j + (1 - \alpha_j^k) u_j, \tag{5.33}$$

with $n = n(k)$,

$$\alpha_j^k = \frac{c_j^k \prod^n (d_i^k)^{P_{ij}}}{1 + c_j^k \prod^n (d_i^k)^{P_{ij}}}, \tag{5.34}$$

$$c_j^k = \frac{(x_j^k - u_j)}{(v_j - x_j^k)}, \tag{5.35}$$

and

$$d_j^k = \frac{(y_i - (Pu)_i)((Pv)_i - (Px^k)_i)}{((Pv)_i - y_i)((Px^k)_i - (Pu)_i)}, \tag{5.36}$$

where $\prod^n$ denotes the product over those indices i in $B_{n(k)}$. Notice that, at each step of the iteration, x_j^k is a convex combination of the endpoints u_j and v_j, so that x_j^k lies in the interval $[u_j, v_j]$.

We have the following theorem concerning the convergence of the ABMART algorithm:

Theorem 5.2 *If there is a solution of the system $Px = y$ that satisfies the constraints $u_j \leq x_j \leq v_j$ for each j, then, for any N and any choice of the blocks B_n, the ABMART sequence converges to that constrained solution of $Px = y$ for which the Fermi-Dirac generalized entropic distance from x to x^0,*

$$KL(x - u, x^0 - u) + KL(v - x, v - x^0),$$

is minimized. If there is no constrained solution of $Px = y$, then, for $N = 1$, the ABMART sequence converges to the minimizer of

$$KL(Px - Pu, y - Pu) + KL(Pv - Px, Pv - y)$$

for which

$$KL(x - u, x^0 - u) + KL(v - x, v - x^0)$$

is minimized.

The proof is similar to that for RBI-SMART and is found in [38].

5.13 The ABEMML Algorithm

We make the same assumptions as in the previous section. The iterative step of the ABEMML algorithm is

$$x_j^{k+1} = \alpha_j^k v_j + (1 - \alpha_j^k) u_j, \tag{5.37}$$

where

$$\alpha_j^k = \gamma_j^k / d_j^k, \tag{5.38}$$

$$\gamma_j^k = (x_j^k - u_j) e_j^k, \tag{5.39}$$

$$\beta_j^k = (v_j - x_j^k) f_j^k, \tag{5.40}$$

$$d_j^k = \gamma_j^k + \beta_j^k, \tag{5.41}$$

$$e_j^k = \left(1 - \sum_{i \in B_n} P_{ij}\right) + \sum_{i \in B_n} P_{ij} \left(\frac{y_i - (Pu)_i}{(Px^k)_i - (Pu)_i}\right), \tag{5.42}$$

and

$$f_j^k = \left(1 - \sum_{i \in B_n} P_{ij}\right) + \sum_{i \in B_n} P_{ij} \left(\frac{(Pv)_i - y_i}{(Pv)_i - (Px^k)_i}\right). \tag{5.43}$$

We have the following theorem concerning the convergence of the ABEMML algorithm:

Theorem 5.3 *If there is a solution of the system $Px = y$ that satisfies the constraints $u_j \leq x_j \leq v_j$ for each j, then, for any N and any choice of the blocks B_n, the ABEMML sequence converges to such a constrained solution of $Px = y$. If there is no constrained solution of $Px = y$, then, for $N = 1$, the ABEMML sequence converges to a constrained minimizer of*

$$KL(y - Pu, Px - Pu) + KL(Pv - y, Pv - Px).$$

The proof is similar to that for RBI-EMML and is to be found in [38]. In contrast to the ABMART theorem, this is all we can say about the limits of the ABEMML sequences.

Open Question: How does the limit of the ABEMML iterative sequence depend, in the consistent case, on the choice of blocks, and, in general, on the choice of x^0?

Chapter 6

Alternating Minimization

6.1 The Alternating-Minimization Framework

As we have seen, the SMART and the EMML algorithms are best derived as alternating minimization (AM) algorithms. The main reference for alternating minimization is the paper [81] of Csiszár and Tusnády. As the authors of [153] remark, the geometric argument in [81] is "deep, though hard to follow". The main reason for the difficulty, I feel, is that the key to their convergence theorem, what they call the *five-point property*, appears to be quite ad hoc and the only good reason for using it that they give is that it works. As we shall see, all AM algorithms can be reformulated as AF methods. When this is done, the five-point property converts precisely into the SUMMA inequality; therefore, all AM methods for which the five-point property of [81] holds fall into the SUMMA class (see [50]).

The alternating minimization (AM) approach provides a useful framework for the derivation of iterative optimization algorithms. In this section we discuss the five-point property of [81] and use it to obtain a somewhat

simpler proof of convergence for their AM algorithm. We then show that all AM algorithms with the five-point property are in the SUMMA class.

6.2 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, $\beta = \inf_{p\in P,\, q\in Q} \Theta(p,q) < +\infty$. We assume also that $\beta > -\infty$; in many applications, the function $\Theta(p,q)$ is nonnegative, so this additional assumption is unnecessary. We do not always assume there are $\hat{p} \in P$ and $\hat{q} \in Q$ such that $\Theta(\hat{p},\hat{q}) = \beta$; 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 $\Theta(p^n, q^n) \downarrow \beta$.

6.3 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 nonnegative, 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 β, since we have

$$\Theta(p^n, q^n) \geq \Theta(p^{n+1}, q^n) \geq \Theta(p^{n+1}, q^{n+1}). \tag{6.1}$$

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

We know two things:

$$\Theta(p^{n+1}, q^n) - \Theta(p^{n+1}, q^{n+1}) \geq 0, \tag{6.2}$$

and

$$\Theta(p^n, q^n) - \Theta(p^{n+1}, q^n) \geq 0. \tag{6.3}$$

Equation 6.3 can be strengthened to

$$\Theta(p, q^n) - \Theta(p^{n+1}, q^n) \geq 0. \tag{6.4}$$

We need to make these inequalities more precise.

6.4 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

$$\Theta(p, q) + \Theta(p, q^{n-1}) \geq \Theta(p, q^n) + \Theta(p^n, q^{n-1}). \tag{6.5}$$

6.5 The Main Theorem for AM

We want to find sufficient conditions for the sequence $\{\Theta(p^n, q^n)\}$ to converge to β, that is, for $\beta^* = \beta$. The following is the main result of [81].

Theorem 6.1 *If the five-point property holds then $\beta^* = \beta$.*

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

$$\Theta(p', q^{n-1}) - \Theta(p^n, q^{n-1}) \geq \Theta(p', q^n) - \Theta(p', q'), \tag{6.6}$$

so that

$$\Theta(p', q^{n-1}) - \Theta(p', q^n) \geq \Theta(p^n, q^{n-1}) - \Theta(p', q') \geq 0. \tag{6.7}$$

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 (6.7) must therefore converge to zero, which is a contradiction. We conclude that $\beta^* = \beta$ whenever the five-point property holds in AM. ∎

6.6 AM as SUMMA

I have not come across any explanation for the five-point property other than it works. I was quite surprised when I discovered that AM algorithms can be reformulated as algorithms minimizing a function $f : P \to \mathbb{R}$ and that the five-point property is then the SUMMA condition in disguise. We show now that the SUMMA class of AF methods includes all the AM algorithms for which the five-point property holds.

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 nth step of AM we minimize

$$G_n(p) = \Theta(p, q^{n-1}) = \Theta(p, q(p)) + \Big(\Theta(p, q^{n-1}) - \Theta(p, q(p))\Big) \tag{6.8}$$

to get p^n. With

$$g_n(p) = \Big(\Theta(p, q^{n-1}) - \Theta(p, q(p))\Big) \geq 0, \tag{6.9}$$

we can write

$$G_n(p) = f(p) + g_n(p). \tag{6.10}$$

According to the five-point property, we have

$$G_n(p) - G_n(p^n) \geq \Theta(p, q^n) - \Theta(p, q(p)) = g_{n+1}(p). \tag{6.11}$$

It follows that AM is a member of the SUMMA class.

6.7 The Three- and Four-Point Properties

In [81] 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

$$\Theta(p, q^n) - \Theta(p^{n+1}, q^n) \geq \Delta(p, p^{n+1}), \tag{6.12}$$

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

The Four-Point Property

$$\Delta(p, p^{n+1}) + \Theta(p, q) \geq \Theta(p, q^{n+1}), \tag{6.13}$$

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

$$\Delta(p, \tilde{p}) = \Theta(p, q(\tilde{p})) - \Theta(p, q(p)), \tag{6.14}$$

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.

6.8 Alternating 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 $\|p - q\|^2$ or $KL(p, q)$. In the case of $\Theta(p, q) = \|p - q\|^2$, 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.

6.9 Bregman Distances

Let $f : \mathbb{R}^J \to \mathbb{R}$ be a Bregman function [26, 72, 31], 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

$$D_f(x, z) = f(x) - f(z) - \langle \nabla f(z), x - z \rangle. \tag{6.15}$$

For example, the KL distance is a Bregman distance with associated Bregman function

$$f(x) = \sum_{j=1}^{J} x_j \log x_j \ - x_j. \tag{6.16}$$

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

$$\langle \nabla f(p^{n+1}) - \nabla f(q^n), p - p^{n+1} \rangle \geq 0, \tag{6.17}$$

for all $p \in P$. Since

$$D_f(p, q^n) - D_f(p^{n+1}, q^n) =$$

$$D_f(p, p^{n+1}) + \langle \nabla f(p^{n+1}) - \nabla f(q^n), p - p^{n+1} \rangle, \tag{6.18}$$

it follows that the three-point property holds, with

$$\Theta(p, q) = D_f(p, q), \tag{6.19}$$

and

$$\Delta(p, \hat{p}) = D_f(p, \tilde{p}). \tag{6.20}$$

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 [10]). Now we can invoke a lemma due to Eggermont and LaRiccia [90].

6.10 The Eggermont-LaRiccia Lemma

Lemma 6.1 *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 ∇_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,

$$\langle \nabla_1 D_f(p^n,q^n), p-p^n\rangle = \langle \nabla f(p^n) - \nabla f(q^n), p-p^n\rangle.$$

It follows that

$$D_f(p,q^n) - D_f(p,p^n) = D_f(p^n,q^n) + \langle \nabla_1 D_f(p^n,q^n), p-p^n\rangle$$

$$\leq D_f(p,q) - \langle \nabla_2 D_f(p^n,q^n), q-q^n\rangle \leq D_f(p,q).$$

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.

6.11 Minimizing a Proximity Function

We present now an example of alternating Bregman distance minimization taken from [41]. The problem is the *convex feasibility problem* (CFP), to find a member of the intersection $C \subseteq \mathbb{R}^J$ of finitely many closed convex sets C_i, $i = 1, ..., I$, or, failing that, to minimize the proximity function

$$F(x) = \sum_{i=1}^{I} D_i(\overleftarrow{P}_i x, x), \tag{6.21}$$

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 $P \subseteq \mathbb{R}^{IJ}$ 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 $Q \subseteq \mathbb{R}^{IJ}$ is the *diagonal* subset, meaning that the elements of Q are the I-fold product of a single x; that is $Q = \{d(x) = (x, x, ..., x) \in \mathbb{R}^{IJ}\}$. We then take

$$\Theta(p, q) = \sum_{i=1}^{I} D_i(c^i, x), \tag{6.22}$$

and $\Delta(p, \tilde{p}) = \Theta(p, \tilde{p})$.

In [64] 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.

6.12 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

$$F(x) = \sum_{i=1}^{I} D_i(\overleftarrow{P}_i x, x) \tag{6.23}$$

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 [12] the authors note that the alternating minimization algorithm of [41] involves right and left Bregman projections, which suggests to them iterative methods involving a wider class of operators that they call "Bregman retractions".

6.13 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.

6.14 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

$$C_i = \{z | (Az)_i = b_i\}, \tag{6.24}$$

and $P_i x$ be the orthogonal projection of x onto C_i. Then

$$(P_i x)_j = x_j + \alpha_i A_{ij}(b_i - (Ax)_i), \tag{6.25}$$

where

$$(\alpha_i)^{-1} = \sum_{j=1}^{J} A_{ij}^2. \tag{6.26}$$

Let

$$F(x) = \sum_{i=1}^{I} \|P_i x - x\|^2. \tag{6.27}$$

Using alternating minimization on this proximity function gives Cimmino's algorithm, with the iterative step

$$x_j^k = x_j^{k-1} + \frac{1}{I}\sum_{i=1}^{I} \alpha_i A_{ij}(b_i - (Ax^{k-1})_i). \tag{6.28}$$

6.15 Simultaneous Projection for Convex Feasibility

Now we let C_i be any closed convex subsets of $\mathbb{R}^J$ and define $F(x)$ as in the previous section. Again, we apply alternating minimization. The iterative step of the resulting algorithm is

$$x^k = \frac{1}{I}\sum_{i=1}^{I} P_i x^{k-1}. \tag{6.29}$$

The objective here is to minimize $F(x)$, if there is a minimum.

6.16 The Bauschke-Combettes-Noll Problem

In [13] Bauschke, Combettes and Noll consider the following problem: minimize the function

$$\Theta(p,q) = \Lambda(p,q) = \phi(p) + \psi(q) + D_f(p,q), \tag{6.30}$$

where ϕ and ψ are convex on $\mathbb{R}^J$, $D = D_f$ is a Bregman distance, and $P = Q$ is the interior of the domain of f. They assume that

$$\beta = \inf_{(p,q)} \Lambda(p,q) > -\infty, \tag{6.31}$$

and seek a sequence $\{(p^n, q^n)\}$ such that $\{\Lambda(p^n, q^n)\}$ converges to β. The sequence is obtained by the AM method, as in our previous discussion. They prove that, if the Bregman distance is jointly convex, then $\{\Lambda(p^n, q^n)\} \downarrow \beta$. In this section 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

$$\Lambda(p, q^{n-1}) - \Lambda(p^n, q^{n-1}) \geq \Lambda(p, q^n) - \Lambda(p,q). \tag{6.32}$$

Lemma 6.2 *The inequality in (6.32) is equivalent to*

$$\Lambda(p,q) - \Lambda(p^n,q^n) \geq$$

$$D(p,q^n) + D(p^n,q^{n-1}) - D(p,q^{n-1}) - D(p^n,q^n). \tag{6.33}$$

Proof: The proof is left to the reader. ∎

By the joint convexity of $D(p,q)$ and the convexity of ϕ and ψ we have

$$\Lambda(p,q) - \Lambda(p^n,q^n) \geq$$

$$\langle \nabla_p \Lambda(p^n,q^n), p - p^n \rangle + \langle \nabla_q \Lambda(p^n,q^n), q - q^n \rangle, \tag{6.34}$$

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

$$\langle \nabla_q \Lambda(p^n,q^n), q - q^n \rangle = 0, \tag{6.35}$$

for all q. Therefore,

$$\Lambda(p,q) - \Lambda(p^n,q^n) \geq \langle \nabla_p \Lambda(p^n,q^n), p - p^n \rangle . \tag{6.36}$$

We have

$$\langle \nabla_p \Lambda(p^n,q^n), p - p^n \rangle =$$

$$\langle \nabla f(p^n) - \nabla f(q^n), p - p^n \rangle + \langle \nabla \phi(p^n), p - p^n \rangle. \tag{6.37}$$

Since p^n minimizes $\Lambda(p,q^{n-1})$, we have

$$\nabla_p \Lambda(p^n,q^{n-1}) = 0, \tag{6.38}$$

or

$$\nabla \phi(p^n) = \nabla f(q^{n-1}) - \nabla f(p^n), \tag{6.39}$$

so that

$$\langle \nabla_p \Lambda(p^n,q^n), p - p^n \rangle = \langle \nabla f(q^{n-1}) - \nabla f(q^n), p - p^n \rangle$$

$$= D(p,q^n) + D(p^n,q^{n-1}) - D(p,q^{n-1}) - D(p^n,q^n). \tag{6.40}$$

Using (6.36) we obtain the inequality in (6.33). 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 β; this is Corollary 4.3 of [13].

If $\psi = 0$, then $\{\Lambda(p^n,q^n)\}$ converges to β, even without the assumption that the distance D_f is jointly convex. In such cases, $\Lambda(p,q)$ has the form of the objective function in proximal minimization and therefore the problem falls into the SUMMA class (see Lemma 3.1).

Chapter 7

The Baillon–Haddad Theorem Revisited

In [14] Heinz Bauschke and Patrick Combettes extend the Baillon–Haddad Theorem to include several additional conditions equivalent to the gradient of a convex function being nonexpansive. These additional conditions involve the Moreau envelope and the Fenchel conjugate. We review these concepts first, and then present their extended Baillon–Haddad Theorem.

7.1 The Fenchel Conjugate

We let $f : \mathcal{H} :\to (-\infty, +\infty]$ be proper. The conjugate of the function f is the function f^* given by

$$f^*(a) = \sup_{x\in\mathcal{H}}\{\langle a, x\rangle - f(x)\}. \tag{7.1}$$

The conjugate of f^* is defined in the obvious way.

7.2 The Moreau Envelope

The Moreau envelope of the function $f : \mathcal{H} \to (-\infty, \infty]$ is the continuous convex function

$$m_f(x) = \mathrm{env}_f(x) = \inf_{y\in C}\{f(y) + \frac{1}{2}\|x - y\|^2\}. \tag{7.2}$$

We assume, from now on, that f is closed, proper, and convex, in which case the infimum is uniquely attained at $y = \text{prox}_f(x)$. Since $y = z$ minimizes $F(y) = f(y) + \frac{1}{2}\|x - y\|^2$ if and only if $0 \in \partial F(z)$, we see, using Equation 1.19 of Proposition 1.4, that

$$0 \in \partial f(\text{prox}_f(x)) - x + \text{prox}_f(x),$$

or

$$x \in \partial f(\text{prox}_f(x)) + \text{prox}_f(x).$$

This characterization of $z = \text{prox}_f(x)$, which we restate now as Proposition 7.1, can be obtained without relying on Proposition 1.4. The proof we give here is that of Proposition 12.26 of [15].

Proposition 7.1 *A point $p \in \mathcal{H}$ is $p = \text{prox}_f(x)$ if and only if $x \in p + \partial f(p)$.*

Proof: Suppose that $p = \text{prox}_f(x)$. Then

$$f(p) + \frac{1}{2}\|x - p\|^2 \leq f(y) + \frac{1}{2}\|x - y\|^2,$$

for all $y \in \mathcal{H}$. Therefore,

$$f(y) - f(p) \geq \frac{1}{2}\|x - p\|^2 - \frac{1}{2}\|x - y\|^2 = \langle y - p, x - p\rangle - \frac{1}{2}\|p - y\|^2.$$

But we want to show that

$$f(y) - f(p) \geq \langle y - p, x - p\rangle.$$

To get this we need to do more work. Let $p_\alpha = (1 - \alpha)p + \alpha y$, for some $\alpha \in (0, 1)$. Then

$$f(p) + \frac{1}{2}\|x - p\|^2 \leq f(p_\alpha) + \frac{1}{2}\|x - p_\alpha\|^2.$$

From

$$f(p_\alpha) \leq (1 - \alpha)f(p) + \alpha f(y),$$

we can get

$$f(y) - f(p) \geq \langle y - p, x - p\rangle - \frac{\alpha}{2}\|y - p\|^2,$$

for all α in $(0, 1)$. Now let α go to zero. ∎

Using Proposition 7.1 we can prove the following Proposition.

Proposition 7.2 *The operator $T = \text{prox}_f$ is firmly nonexpansive.*

Proof: Let $p = \text{prox}_f(x)$ and $q = \text{prox}_f(y)$. Then we have $x - p \in \partial f(p)$, and $y - q \in \partial f(q)$. Therefore,

$$f(q) - f(p) \geq \langle x - p, q - p \rangle,$$

and

$$f(p) - f(q) \geq \langle y - q, p - q \rangle,$$

from which we obtain

$$\langle p - q, x - y \rangle \geq \|p - q\|^2.$$

∎

If $C \subseteq \mathcal{H}$ is a nonempty, closed, convex subset, and $f = \iota_C$, then $\text{prox}_f = P_C$. Also $\partial f(x) = N_C(x)$, the normal cone to C at x and

$$J_{\partial \iota_C} = J_{N_C} = \text{prox}_{\iota_C} = P_C.$$

Proposition 7.3 *The Moreau envelope $m_f(x) = \text{env}_f(x)$ is Fréchet differentiable and*

$$\nabla m_f(x) = x - \text{prox}_f(x). \tag{7.3}$$

Proof: Let $p = \text{prox}_f(x)$ and $q = \text{prox}_f(y)$. Then

$$m_f(y) - m_f(x) = f(q) + \frac{1}{2}\|q - y\|^2 - f(p) - \frac{1}{2}\|p - x\|^2$$

$$= f(q) - f(p) + \frac{1}{2}\|q - y\|^2 - \frac{1}{2}\|p - x\|^2 \geq \langle q - p, x - p \rangle + \frac{1}{2}\|q - y\|^2 - \frac{1}{2}\|p - x\|^2.$$

Consequently,

$$m_f(y) - m_f(x) \geq \frac{1}{2}\|y - q - x + p\|^2 + \langle y - x, x - p \rangle \geq \langle y - x, x - p \rangle.$$

Similarly,

$$m_f(x) - m_f(y) \geq \langle x - y, y - q \rangle.$$

Then

$$0 \leq m_f(y) - m_f(x) - \langle y - x, x - p \rangle \leq \|x - y\|^2 - \|q - p\|^2,$$

where we have used the fact that $T = \text{prox}_f$ is fne. Therefore,

$$\frac{m_f(y) - m_f(x) - \langle y - x, x - p \rangle}{\|x - y\|} \leq \|x - y\|.$$

Now let $y \to x$. ∎

Proposition 7.4 *Let $f : \mathcal{H} \to (-\infty, \infty]$ be closed, proper, and convex. Then $x = z$ minimizes $f(x)$ if and only if $x = z$ minimizes $m_f(x)$.*

Proof: We have $x = z$ minimizes $m_f(x)$ if and only if

$$0 = \nabla m_f(z) = z - \text{prox}_f(z),$$

or

$$\text{prox}_f(z) = z.$$

Since, for any x, $z = \text{prox}_f(x)$ if and only if $x - z \in \partial f(z)$, it follows, using $x = z$, that $0 \in \partial f(z)$ and so z minimizes f, ∎

7.3 Infimal Convolution

Let $f : \mathcal{H} \to \mathbb{R}$ and $g : \mathcal{H} \to \mathbb{R}$ be arbitrary. Then the *infimal convolution* of f and g, written $f \oplus g$, is

$$(f \oplus g)(x) = \inf_y \{f(y) + g(x - y)\}; \tag{7.4}$$

see Lucet [125] for details. Using $g(x) = q(x) = \frac{1}{2}\|x\|^2$, we have $f \oplus q = m_f$.

Proposition 7.5 *Let f and g be functions from $\mathcal{H}$ to $\mathbb{R}$. Then we have $(f \oplus g)^* = f^* + g^*$.*

Proof: Select $a \in \mathcal{H}$. Then

$$(f \oplus g)^*(a) = \sup_x \left(\langle a, x \rangle - \inf_y \{f(y) + g(x - y)\} \right)$$

$$= \sup_y \left(\langle y, a \rangle - f(y) + \sup_x \{\langle x - y, a \rangle - g(x - y)\} \right) = f^*(a) + g^*(a).$$

∎

Corollary 7.1 *With $q(x) = \frac{1}{2}\|x\|^2 = q^*(x)$ in place of $g(x)$, we have*

1. $(m_f)^* = (f \oplus q)^* = f^* + q$;
2. $m_f = f \oplus q = (f^* + q)^*$; *and*
3. $m_{f^*} = f^* \oplus q = (f + q)^*$.

Proposition 7.6 *Let $f : \mathcal{H} \to \mathbb{R}$ be closed and convex. The following hold:*

1. $m_f = q - (f+q)^*$;
2. $m_f + m_{f^*} = q$; *and*
3. $\operatorname{prox}_f + \operatorname{prox}_{f^*} = I$.

Proof: First we prove 1. For any $x \in \mathcal{H}$ we have

$$m_f(x) = \inf_y\{f(y) + q(x-y)\} = \inf_y\{f(y) + q(x) + q(y) - \langle x, y\rangle\}$$
$$= q(x) - \sup_y\{\langle x, y\rangle - f(y) - q(y)\} = q(x) - (f+q)^*(x).$$

Assertion 2. then follows from the previous corollary, and we get Assertion 3. by taking gradients. ∎

Proposition 7.7 *Let $f : \mathcal{H} \to \mathbb{R}$ be closed and convex, $q(x) = \frac{1}{2}\|x\|^2$, $g(x) = q(x) - f(x)$, and $h(x) = f^*(x) - q(x)$. If g is convex, then so is h.*

Proof: We have

$$f(x) = q(x) - g(x) = q(x) - g^{**}(x) = q(x) - \sup_u\{\langle u, x\rangle - g^*(u)\}$$
$$= \inf_u\{q(x) - \langle u, x\rangle + g^*(u)\}.$$

Therefore

$$f^*(a) = \sup_x \sup_u\{\langle a, x\rangle + \langle u, x\rangle - q(x) - g^*(u)\}$$

so

$$f^*(a) = \sup_u\{q^*(a+u) - g^*(u)\}.$$

From

$$q^*(a+u) = \frac{1}{2}\|a+u\|^2 = \frac{1}{2}\|a\|^2 + \langle a, u\rangle + \frac{1}{2}\|u\|^2,$$

we get

$$f^*(a) = \frac{1}{2}\|a\|^2 + (g^* - q)^*(a),$$

or

$$h(a) = f^*(a) - \frac{1}{2}\|a\|^2 = (g^* - q)(a) = \sup_x\{\langle a, x\rangle - g^*(x) + q(x)\},$$

which is the supremum of a family of affine functions in the variable a, and so is convex. ∎

Proposition 7.8 *Let $f : \mathcal{H} \to \mathbb{R}$ be closed and convex, $q(x) = \frac{1}{2}\|x\|^2$, and $h(x) = f^*(x) - q(x)$. If h is convex, then $f = m_{h^*}$.*

Proof: From $h = f^* - q$ we get $f^* = h + q$, so that

$$f = f^{**} = (h+q)^* = h^* \oplus q = m_{h^*}.$$

∎

7.4 The Extended Baillon–Haddad Theorem

Now we are in a position to consider the extended Baillon–Haddad Theorem of Bauschke and Combettes. To avoid technicalities, we present a slightly simplified version of the theorem in [14, 15].

Theorem 7.1 *Let $f : \mathcal{H} \to \mathbb{R}$ be closed and convex, $q(x) = \frac{1}{2}\|x\|^2$, $g(x) = q(x) - f(x)$, and $h(x) = f^*(x) - q(x)$. The following are equivalent:*

1. *f is Fréchet differentiable and the operator $T = \nabla f$ is nonexpansive;*
2. *g is convex;*
3. *h is convex;*
4. *$f = m_{h^*}$;*
5. *$\nabla f = \text{prox}_h = I - \text{prox}_{h^*}$;*
6. *f is Fréchet differentiable and the operator $T = \nabla f$ is firmly nonexpansive.*

Proof: Showing 1. implies 2. was done previously, in the earlier version of the Baillon–Haddad Theorem. To show that 2. implies 3. use Proposition 7.7. Assuming 3., we get 4. using Proposition 7.8. Then to get 4. implies 5. we use Proposition 7.3 and Proposition 7.6. Finally, we assume 5. and get 6. from Proposition 7.2 and the continuity of ∇f. ∎

As the authors of [14] noted, their proof was new and shorter than those found in the literature up to that time, since several of the equivalences they employ were already established by others. The equivalence of conditions 2., 3., and 4. was established in [133]. The equivalence of conditions 1., 3., 4., and 6. was shown in Euclidean spaces in [144], Proposition 12.60, using different techniques.

Chapter 8

Appendix: Bregman–Legendre Functions

In [9] Bauschke and Borwein show convincingly that the Bregman–Legendre functions provide the proper context for the discussion of Bregman projections onto closed convex sets in $\mathbb{R}^J$. The summary here follows closely the discussion given in [9].

8.1 Essential Smoothness and Essential Strict Convexity

Let $f : \mathbb{R}^J \to (-\infty, \infty]$ be closed, proper and convex, with essential domain $D = \{x | f(x) \in \mathbb{R}\}$. Following [143] we say that f is *essentially smooth* if int$D$ is not empty, f is differentiable on int$D$ and $x^n \in \text{int}D$, with $x^n \to x \in \text{bd}D$, implies that $\|\nabla f(x^n)\|_2 \to +\infty$. Here int$D$ and bd$D$ denote the interior and boundary of the set D. A closed proper convex function f is *essentially strictly convex* if f is strictly convex on every convex subset of dom $\partial f = \{x | \partial f(x) \neq \emptyset\}$.

The closed proper convex function f is essentially smooth if and only if the subdifferential $\partial f(x)$ is empty for $x \in \text{bd}D$ and is $\{\nabla f(x)\}$ for $x \in \text{int}D$ (so f is differentiable on int$D$) if and only if the function f^* is essentially strictly convex.

Definition 8.1 *A closed proper convex function f is said to be a* Legendre function *if it is both essentially smooth and essentialy strictly convex.*

So f is Legendre if and only if its conjugate function is Legendre, in which case the gradient operator ∇f is a topological isomorphism with

∇f^* as its inverse. The gradient operator ∇f maps int dom f onto int dom f^*. If int dom $f^* = \mathbb{R}^J$ then the range of ∇f is $\mathbb{R}^J$ and the equation $\nabla f(x) = y$ can be solved for every $y \in \mathbb{R}^J$. In order for int dom $f^* = \mathbb{R}^J$ it is necessary and sufficient that the Legendre function f be *super-coercive*, that is,

$$\lim_{||x||_2 \to +\infty} \frac{f(x)}{||x||_2} = +\infty. \tag{8.1}$$

If the effective domain of f is bounded, then f is super-coercive and its gradient operator is a mapping onto the space $\mathbb{R}^J$.

8.2 Bregman Projections onto Closed Convex Sets

Let f be a closed proper convex function that is differentiable on the nonempty set int$D$. The corresponding *Bregman distance* $D_f(x, z)$ is defined for $x \in \mathbb{R}^J$ and $z \in$ int$D$ by

$$D_f(x, z) = f(x) - f(z) - \langle \nabla f(z), x - z \rangle. \tag{8.2}$$

Note that $D_f(x, z) \geq 0$ always and that $D_f(x, z) = +\infty$ is possible. If f is essentially strictly convex then $D_f(x, z) = 0$ implies that $x = z$.

Let K be a nonempty closed convex set with $K \cap$ int$D \neq \emptyset$. Pick $z \in$ int$D$. The *Bregman projection* of z onto K, with respect to f, is

$$P_K^f(z) = \text{argmin}_{x \in K \cap D} D_f(x, z). \tag{8.3}$$

If f is essentially strictly convex, then $P_K^f(z)$ exists. If f is strictly convex on D then $P_K^f(z)$ is unique. If f is Legendre, then $P_K^f(z)$ is uniquely defined and is in int$D$; this last condition is sometimes called *zone consistency*.

Example: Let $J = 2$ and $f(x)$ be the function that is equal to one-half the norm squared on D, the nonnegative quadrant, $+\infty$ elsewhere. Let K be the set $K = \{(x_1, x_2) | x_1 + x_2 = 1\}$. The Bregman projection of $(2, 1)$ onto K is $(1, 0)$, which is not in int$D$. The function f is not essentially smooth, although it is essentially strictly convex. Its conjugate is the function f^* that is equal to one-half the norm squared on D and equal to zero elsewhere; it is essentially smooth, but not essentially strictly convex.

If f is Legendre, then $P_K^f(z)$ is the unique member of $K \cap$ int$D$ satisfying the inequality

$$\langle \nabla f(P_K^f(z)) - \nabla f(z), P_K^f(z) - c \rangle \geq 0, \tag{8.4}$$

for all $c \in K$. From this we obtain the *Bregman Inequality*:

$$D_f(c,z) \geq D_f(c, P_K^f(z)) + D_f(P_K^f(z), z), \tag{8.5}$$

for all $c \in K$.

8.3 Bregman–Legendre Functions

Following Bauschke and Borwein [9], we say that a Legendre function f is a *Bregman–Legendre* function if the following properties hold:

B1: for x in D and any $a > 0$ the set $\{z | D_f(x,z) \leq a\}$ is bounded.
B2: if x is in D but not in int$D$, for each positive integer n, y^n is in int$D$ with $y^n \to y \in \text{bd}D$ and if $\{D_f(x,y^n)\}$ remains bounded, then $D_f(y,y^n) \to 0$, so that $y \in D$.
B3: if x^n and y^n are in int$D$, with $x^n \to x$ and $y^n \to y$, where x and y are in D but not in int$D$, and if $D_f(x^n,y^n) \to 0$ then $x = y$.

Bauschke and Borwein then prove that Bregman's SGP method converges to a member of K provided that one of the following holds: 1) f is Bregman–Legendre; 2) $K \cap \text{int}D \neq \emptyset$ and dom f^* is open; or 3) dom f and dom f^* are both open.

The Bregman functions form a class closely related to the Bregman-Legendre functions. For details see [31].

8.3.1 Useful Results about Bregman–Legendre Functions

The following results are proved in somewhat more generality in [9].
R1: If $y^n \in \text{int dom } f$ and $y^n \to y \in \text{int dom } f$, then $D_f(y,y^n) \to 0$.
R2: If x and $y^n \in \text{int dom } f$ and $y^n \to y \in \text{bd dom } f$, then $D_f(x,y^n) \to +\infty$.
R3: If $x^n \in D$, $x^n \to x \in D$, $y^n \in \text{int } D$, $y^n \to y \in D$, $\{x,y\} \cap \text{int } D \neq \emptyset$ and $D_f(x^n,y^n) \to 0$, then $x = y$ and $y \in \text{int } D$.
R4: If x and y are in D, but are not in int D, $y^n \in \text{int } D$, $y^n \to y$ and $D_f(x,y^n) \to 0$, then $x = y$.
As a consequence of these results we have the following.
R5: If $\{D_f(x,y^n)\} \to 0$, for $y^n \in \text{int } D$ and $x \in \mathbb{R}^J$, then $\{y^n\} \to x$.

Proof of R5: Since $\{D_f(x,y^n)\}$ is eventually finite, we have $x \in D$. By Property B1 above it follows that the sequence $\{y^n\}$ is bounded; without loss of generality, we assume that $\{y^n\} \to y$, for some $y \in \overline{D}$. If x is in int

D, then, by result R2 above, we know that y is also in int D. Applying result R3, with $x^n = x$, for all n, we conclude that $x = y$. If, on the other hand, x is in D, but not in int D, then y is in D, by result R2. There are two cases to consider: 1) y is in int D; 2) y is not in int D. In case 1) we have $D_f(x, y^n) \to D_f(x, y) = 0$, from which it follows that $x = y$. In case 2) we apply result R4 to conclude that $x = y$. ∎

Bibliography

[1] Anderson, A. and Kak, A. (1984) "Simultaneous algebraic reconstruction technique (SART): a superior implementation of the ART algorithm." *Ultrasonic Imaging*, **6** pp. 81–94.

[2] 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.

[3] Axelsson, O. (1994) *Iterative Solution Methods*. Cambridge, UK: Cambridge University Press.

[4] Baillon, J.-B., and Haddad, G. (1977) "Quelques propriétés des opérateurs angle-bornés et n-cycliquement monotones." *Israel J. of Mathematics*, **26**, pp. 137-150.

[5] Baillon, J.-B., Bruck, R.E., and Reich, S. (1978) "On the asymptotic behavior of nonexpansive mappings and semigroups in Banach spaces." *Houston Journal of Mathematics*, **4**, pp. 1–9.

[6] Bauschke, H. (1996) "The approximation of fixed points of compositions of nonexpansive mappings in Hilbert space." *Journal of Mathematical Analysis and Applications*, **202**, pp. 150–159.

[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. (1997) "Legendre functions and the method of random Bregman projections." *Journal of Convex Analysis*, **4**, pp. 27–67.

[10] Bauschke, H., and Borwein, J. (2001) "Joint and separate convexity of the Bregman distance." in [30], pp. 23–36.

[11] Bauschke, H., and Combettes, P. (2001) "A weak-to-strong convergence principle for Fejér monotone methods in Hilbert spaces." *Mathematics of Operations Research*, **26**, pp. 248–264.

[12] Bauschke, H., and Combettes, P. (2003) "Iterating Bregman retractions." *SIAM Journal on Optimization*, **13**, pp. 1159–1173.

[13] Bauschke, H., Combettes, P., and Noll, D. (2006) "Joint minimization with alternating Bregman proximity operators." *Pacific Journal of Optimization*, **2**, pp. 401–424.

[14] Bauschke, H., and Combettes, P. (2010) "The Baillon-Haddad Theorem Revisited." *J. Convex Analysis*, **17**, pp. 781–787.

[15] Bauschke, H., and Combettes, P. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, New York: Springer CMS Books in Mathematics, 2011.

[16] Bauschke, H., and Lewis, A. (2000) "Dykstra's algorithm with Bregman projections: a convergence proof." *Optimization*, **48**, pp. 409–427.

[17] Bauschke, H., Burachik, R., Combettes, P., Elser, V., Luke, D., and Wolkowitz, H., eds. *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, New York: Springer-Verlag, 2011.

[18] Becker, M., Yang, I., and Lange, K. (1997) "EM algorithms without missing data." *Stat. Methods Med. Res.*, **6**, pp. 38–54.

[19] Berinde, V. (2007) *Iterative Approximation of Fixed Points*, Berlin: Springer-Verlag.

[20] Bertero, M., and Boccacci, P. (1998) *Introduction to Inverse Problems in Imaging*, Bristol, UK: Institute of Physics Publishing.

[21] Bertsekas, D.P. (1997) "A new class of incremental gradient methods for least squares problems." *SIAM J. Optim.*, **7**, pp. 913-926.

[22] Bertsekas, D., and Tsitsiklis, J. (1989) *Parallel and Distributed Computation: Numerical Methods*. New Jersey: Prentice-Hall.

[23] Bertsekas, D. *Convex Analysis and Optimization*, Nashua, NH: Athena Scientific, 2003.

[24] Borwein, J. and Lewis, A. (2000) *Convex Analysis and Nonlinear Optimization*. Canadian Mathematical Society Books in Mathematics, New York: Springer-Verlag.

[25] Boyd, S., and Vandenberghe, L. (2004) *Convex Optimization*. Cambridge, England: Cambridge University Press.

[26] 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.

[27] Bregman, L., Censor, Y., and Reich, S. (1999) "Dykstra's algorithm as the nonlinear extension of Bregman's optimization method." *Journal of Convex Analysis*, **6 (2)**, pp. 319–333.

[28] 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.

[29] Bruck, R.E., and Reich, S. (1977) "Nonexpansive projections and resolvents of accretive operators in Banach spaces." *Houston Journal of Mathematics*, **3**, pp. 459–470.

[30] Butnariu, D., Censor, Y., and Reich, S. (eds.) (2001) *Inherently Parallel Algorithms in Feasibility and Optimization and their Applications*, Studies in Computational Mathematics 8. Amsterdam: Elsevier Publ.

[31] Butnariu, D., Byrne, C., and Censor, Y. (2003) "Redundant axioms in the definition of Bregman functions." *Journal of Convex Analysis*, **10**, pp. 245–254.

[32] Byrne, C. (1993) "Iterative image reconstruction algorithms based on cross-entropy minimization." *IEEE Transactions on Image Processing* **IP-2**, pp. 96–103.

[33] 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.

[34] 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.

[35] Byrne, C. (1996) "Block-iterative methods for image reconstruction from projections." *IEEE Transactions on Image Processing* **IP-5**, pp. 792–794.

[36] Byrne, C. (1997) "Convergent block-iterative algorithms for image reconstruction from inconsistent data." *IEEE Transactions on Image Processing* **IP-6**, pp. 1296–1304.

[37] 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.

[38] Byrne, C. (1998) "Iterative algorithms for deblurring and deconvolution with constraints." *Inverse Problems*, **14**, pp. 1455–1467.

[39] Byrne, C. (2000) "Block-iterative interior point optimization methods for image reconstruction from limited data." *Inverse Problems* **16**, pp. 1405–1419.

[40] Byrne, C. (2001) "Bregman-Legendre multi-distance projection algorithms for convex feasibility and optimization." in [30], pp. 87–100.

[41] 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.

[42] Byrne, C. (2002) "Iterative oblique projection onto convex sets and the split feasibility problem." *Inverse Problems* **18**, pp. 441–453.

[43] Byrne, C. (2004) "A unified treatment of some iterative algorithms in signal processing and image reconstruction." *Inverse Problems* **20**, pp. 103–120.

[44] Byrne, C. (2005) "Choosing parameters in block-iterative or ordered-subset reconstruction algorithms." *IEEE Transactions on Image Processing*, **14 (3)**, pp. 321–327.

[45] Byrne, C. (2005) *Signal Processing: A Mathematical Approach*, AK Peters, Publ., Wellesley, MA.

[46] Byrne, C. (2007) *Applied Iterative Methods*, AK Peters, Publ., Wellesley, MA.

[47] Byrne, C. (2008) "Sequential unconstrained minimization algorithms for constrained optimization." *Inverse Problems*, **24(1)**, article no. 015013.

[48] Byrne, C. (2009) "Block-iterative algorithms." *International Transactions in Operations Research*, **16(4)**, pp. 427–463.

[49] 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.

[50] Byrne, C. (2013) "Alternating minimization as sequential unconstrained minimization: a survey." *Journal of Optimization Theory and Applications*, electronic **154(3)**, DOI 10.1007/s1090134-2, (2012), and hardcopy **156(3)**, February, 2013, pp. 554–566.

[51] Byrne, C. (2014) "An elementary proof of convergence of the forward-backward splitting algorithm." *Journal of Nonlinear and Convex Analysis* **15(4)**, pp. 681–691.

[52] Byrne, C. (2014) "On a generalized Baillon–Haddad Theorem for convex functions on Hilbert space." to appear in the *Journal of Convex Analysis*.

[53] Byrne, C., and Eggermont, P. (2011) "EM Algorithms." in *Handbook of Mathematical Methods in Imaging*, Otmar Scherzer, ed., Springer-Science.

[54] Byrne, C., Censor, Y., A. Gibali, A., and Reich, S. (2012) "The split common null point problem." *Journal of Nonlinear and Convex Analysis*, **13**, pp. 759–775.

[55] Byrne, C. (2014) *Iterative Optimization in Inverse Problems*. Boca Raton, FL: CRC Press.

[56] Byrne, C. (2014) *A First Course in Optimization*. Boca Raton, FL: CRC Press.

[57] Cegielski, A. (2010) "Generalized relaxations of nonexpansive operators and convex feasibility problems." *Contemp. Math.*, **513**, pp. 111–123.

[58] Cegielski, A. (2012) *Iterative Methods for Fixed Point Problems in Hilbert Space*. Heidelberg: Springer Lecture Notes in Mathematics 2057.

[59] Cegielski, A., and Censor, Y. (2011) "Opial-type theorems and the common fixed-point problem." in *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, H. Bauschke, R. Burachik, P. Combettes, V. Elser, D. Luke, and H. Wolkowicz (eds.), Springer Optimization and its Applications, Vol. 49, New York: Springer.

[60] Cegielski, A., and Zalas, R. (2013) "Methods for the variational inequality problem over the intersection of fixed points of quasi-nonexpansive operators." *Numer. Funct. Anal. Optimiz.*, **34**, pp. 255–283.

[61] Censor, Y. (1981) "Row-action methods for huge and sparse systems and their applications." *SIAM Review*, **23**, pp. 444–464.

[62] 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.

[63] Censor, Y., Eggermont, P.P.B., and Gordon, D. (1983) "Strong underrelaxation in Kaczmarz's method for inconsistent systems." *Numerische Mathematik* **41**, pp. 83–92.

[64] Censor, Y. and Elfving, T. (1994) "A multi-projection algorithm using Bregman projections in a product space." *Numerical Algorithms*, **8** 221–239.

[65] Censor, Y., Elfving, T., Herman, G.T., and Nikazad, T. (2008) "On diagonally-relaxed orthogonal projection methods." *SIAM Journal on Scientific Computation*, **30(1)**, pp. 473–504.

[66] 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.

[67] Censor, Y., Iusem, A., and Zenios, S. (1998) "An interior point method with Bregman functions for the variational inequality problem with paramonotone operators." *Mathematical Programming*, **81**, pp. 373–400.

[68] Censor, Y., and Reich, S. (1996) "Iterations of paracontractions and firmly nonexpansive operators with applications to feasibility and optimization." *Optimization*, **37**, pp. 323–339.

[69] Censor, Y., and Reich, S. (1998) "The Dykstra algorithm for Bregman projections." *Communications in Applied Analysis*, **2**, pp. 323–339.

[70] Censor, Y. and Segman, J. (1987) "On block-iterative maximization." *J. of Information and Optimization Sciences* **8**, pp. 275–291.

[71] Censor, Y., and Zenios, S.A. (1992) "Proximal minimization algorithm with D-functions." *Journal of Optimization Theory and Applications*, **73(3)**, pp. 451–464.

[72] Censor, Y. and Zenios, S.A. (1997) *Parallel Optimization: Theory, Algorithms and Applications*. New York: Oxford University Press.

[73] Cheney, W., and Goldstein, A. (1959) "Proximity maps for convex sets." *Proc. Amer. Math. Soc.*, **10**, pp. 448–450.

[74] Chidume, Ch. (2009) *Geometric Properties of Banach Spaces and Nonlinear Iterations*. London: Springer.

[75] Cimmino, G. (1938) "Calcolo approssimato per soluzioni dei sistemi di equazioni lineari." *La Ricerca Scientifica XVI, Series II, Anno IX* **1**, pp. 326–333.

[76] Combettes, P. (2000) "Fejér monotonicity in convex optimization."in *Encyclopedia of Optimization*, C.A. Floudas and P. M. Pardalos, editors, Boston: Kluwer Publ.

[77] Combettes, P. (2001) "Quasi-Fejérian analysis of some optimization algorithms." 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.

[78] Combettes, P., and Wajs, V. (2005) "Signal recovery by proximal forward-backward splitting." *Multiscale Modeling and Simulation*, **4(4)**, pp. 1168–1200.

[79] Csiszár, I. (1975) "I-divergence geometry of probability distributions and minimization problems." *The Annals of Probability* **3(1)**, pp. 146–158.

[80] Csiszár, I. (1989) "A geometric interpretation of Darroch and Ratcliff's generalized iterative scaling." *The Annals of Statistics* **17(3)**, pp. 1409–1413.

[81] Csiszár, I. and Tusnády, G. (1984) "Information geometry and alternating minimization procedures." *Statistics and Decisions* **Supp. 1**, pp. 205–237.

[82] Darroch, J. and Ratcliff, D. (1972) "Generalized iterative scaling for log-linear models." *Annals of Mathematical Statistics* **43**, pp. 1470–1480.

[83] Dax, A. (1990) "The convergence of linear stationary iterative processes for solving singular unstructured systems of linear equations." *SIAM Review*, **32**, pp. 611–635.

[84] 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.

[85] De Pierro, A. and Iusem, A. (1990) "On the asymptotic behavior of some alternate smoothing series expansion iterative methods." *Linear Algebra and its Applications* **130**, pp. 3–24.

[86] Deutsch, F., and Yamada, I. (1998) "Minimizing certain convex functions over the intersection of the fixed point sets of non-expansive mappings." *Numerical Functional Analysis and Optimization*, **19**, pp. 33–56.

[87] Dugundji, J. (1970) *Topology* Boston: Allyn and Bacon, Inc.

[88] Dykstra, R. (1983) "An algorithm for restricted least squares regression." *J. Amer. Statist. Assoc.*, **78 (384)**, pp. 837–842.

[89] Eggermont, P.P.B., Herman, G.T., and Lent, A. (1981) "Iterative algorithms for large partitioned linear systems, with applications to image reconstruction." *Linear Algebra and its Applications* **40**, pp. 37–67.

[90] Eggermont, P., and LaRiccia, V. (2001) *Maximum Penalized Likelihood Estimation*. New York: Springer.

[91] Elsner, L., Koltracht, L., and Neumann, M. (1992) "Convergence of sequential and asynchronous nonlinear paracontractions." *Numerische Mathematik*, **62**, pp. 305–319.

[92] Everitt, B., and Hand, D. (1981) *Finite Mixture Distributions* London: Chapman and Hall.

[93] Fessler, J., Ficaro, E., Clinthorne, N., and Lange, K. (1997) "Grouped-coordinate ascent algorithms for penalized-likelihood transmission image reconstruction." *IEEE Trans. Med. Imag.* **16 (2)** pp. 166–175.

[94] Fiacco, A., and McCormick, G. (1990) *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*. Philadelphia, PA: SIAM Classics in Mathematics (reissue).

[95] Fiddy, M. (2008) *private communication*.

[96] Gill, P., Murray, W., and Wright, M. (1981) *Practical Optimization*, Academic Press, San Diego.

[97] 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.

[98] Goebel, K., and Reich, S. (1984) *Uniform Convexity, Hyperbolic Geometry, and Nonexpansive Mappings*, New York: Dekker.

[99] Goldstein, S., and Osher, S. (2008) "The split Bregman algorithm for L^1 regularized problems. " UCLA CAM Report 08-29, UCLA, Los Angeles.

[100] Golshtein, E., and Tretyakov, N. (1996) *Modified Lagrangians and Monotone Maps in Optimization*. New York: John Wiley and Sons, Inc.

[101] 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.

[102] Gubin, L.G., Polyak, B.T. and Raik, E.V. (1967) "The method of projections for finding the common point of convex sets." *USSR Computational Mathematics and Mathematical Physics*, **7**: 1–24.

[103] Hager, W. (1988) *Applied Numerical Linear Algebra*, Englewood Cliffs, NJ: Prentice Hall.

[104] Herman, G. T. (1999) *private communication.*

[105] Herman, G. T. and Meyer, L. (1993) "Algebraic reconstruction techniques can be made computationally efficient." *IEEE Transactions on Medical Imaging* **12**, pp. 600–609.

[106] Hildreth, C. (1957) "A quadratic programming procedure." *Naval Research Logistics Quarterly*, **4**, pp. 79–85. Erratum, ibid., p. 361.

[107] Hiriart-Urruty, J.-B., and Lemaréchal, C. (2001) *Fundamentals of Convex Analysis*. Berlin: Springer.

[108] Hirstoaga, S.A. (2006) "Iterative selection methods for common fixed point problems." J. Math. Anal. Appl., **324**, pp. 1020–1035.

[109] R. Hogg, J. McKean, and A. Craig, *Introduction to Mathematical Statistics*, 6th edition, Prentice Hall (2004).

[110] 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.

[111] Jiang, M., and Wang, G. (2003) "Convergence studies on iterative algorithms for image reconstruction." *IEEE Transactions on Medical Imaging*, **22(5)**, pp. 569–579.

[112] Kaczmarz, S. (1937) "Angenäherte Auflösung von Systemen linearer Gleichungen." *Bulletin de l'Academie Polonaise des Sciences et Lettres* **A35**, pp. 355–357.

[113] Karmarkar, N. (1984) "A new polynomial-time algorithm for linear programming." *Combinatorica*, **4**, pp. 373–395.

[114] Kelley, C.T. (1999) *Iterative Methods for Optimization*, Frontiers in Applied Mathematics, Philadelphia: SIAM Publications.

[115] Korpelevich, G. (1976) "The extragradient method for finding saddle points and other problems." *Ekonomika i Matematcheskie Metody* (in Russian), **12**, pp. 747–756.

[116] Krasnosel'skii, M. (1955) "Two observations on the method of sequential approximations." *Uspeki Mathematicheskoi Nauki* (in Russian), **10(1)**.

[117] Kullback, S. and Leibler, R. (1951) "On information and sufficiency." *Annals of Mathematical Statistics* **22**, pp. 79–86.

[118] Landweber, L. (1951) "An iterative formula for Fredholm integral equations of the first kind." *Amer. J. of Math.* **73**, pp. 615–624.

[119] Lange, K. and Carson, R. (1984) "EM reconstruction algorithms for emission and transmission tomography." *Journal of Computer Assisted Tomography* **8**, pp. 306–316.

[120] 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.

[121] Lange, K., Hunter, D., and Yang, I. (2000) "Optimization transfer using surrogate objective functions (with discussion)." *J. Comput. Graph. Statist.*, **9**, pp. 1–20.

[122] 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.

[123] Lent, A., and Censor, Y. (1980) "Extensions of Hildreth's row-action method for quadratic programming." *SIAM Journal on Control and Optimization*, **18**, pp. 444–454.

[124] Levy, A. (2009) *The Basics of Practical Optimization*. Philadelphia: SIAM Publications.

[125] Lucet, Y. (2010) "What shape is your conjugate? A survey of computational convex analysis and its applications." *SIAM Review*, **52(3)**, pp. 505–542.

[126] Luenberger, D. (1969) *Optimization by Vector Space Methods*. New York: John Wiley and Sons, Inc.

[127] Luo, Z., Ma, W., So, A., Ye, Y., and Zhang, S. (2010) "Semidefinite relaxation of quadratic optimization problems." *IEEE Signal Processing Magazine*, **27 (3)**, pp. 20–34.

[128] Mann, W. (1953) "Mean value methods in iteration." *Proc. Amer. Math. Soc.* **4**, pp. 506–510.

[129] Marzetta, T. (2003) "Reflection coefficient (Schur parameter) representation for convex compact sets in the plane." *IEEE Transactions on Signal Processing*, **51 (5)**, pp. 1196–1210.

[130] McLachlan, G.J. and Krishnan, T. (1997) *The EM Algorithm and Extensions*. New York: John Wiley and Sons, Inc.

[131] 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.

[132] Moreau, J.-J. (1963) "Propriétés des applications 'prox'." *C.R. Acad. Sci. Paris Sér. A Math.*, **256**, pp. 1069–1071.

[133] Moreau, J.-J. (1965) "Proximité et dualité dans un espace hilbertien." *Bull. Soc. Math. France*, **93**, pp. 273–299.

[134] 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.

[135] Nash, S. and Sofer, A. (1996) *Linear and Nonlinear Programming*. New York: McGraw-Hill.

[136] Nesterov, Y., and Nemirovski, A. (1994) *Interior-Point Polynomial Algorithms in Convex Programming*. Philadelphia, PA: SIAM Studies in Applied Mathematics.

[137] Opial, Z. (1967) "Weak convergence of the sequence of successive approximations for nonexpansive mappings." *Bulletin of the American Mathematical Society*, **73**, pp. 591–597.

[138] Ortega, J., and Rheinboldt, W. (2000) *Iterative Solution of Nonlinear Equations in Several Variables*, Classics in Applied Mathematics, 30. Philadelphia, PA: SIAM, 2000.

[139] Reich, S. (1979) "Weak convergence theorems for nonexpansive mappings in Banach spaces." *Journal of Mathematical Analysis and Applications*, **67**, pp. 274–276.

[140] Reich, S. (1980) "Strong convergence theorems for resolvents of accretive operators in Banach spaces." *Journal of Mathematical Analysis and Applications*, pp. 287–292.

[141] Reich, S. (1996) "A weak convergence theorem for the alternating method with Bregman distances." *Theory and Applications of Nonlinear Operators*, New York: Dekker.

[142] Renegar, J. (2001) *A Mathematical View of Interior-Point Methods in Convex Optimization*. Philadelphia, PA: SIAM (MPS-SIAM Series on Optimization).

[143] Rockafellar, R. (1970) *Convex Analysis*. Princeton, NJ: Princeton University Press.

[144] Rockafellar, R.T. and Wets, R. J-B. (2009) *Variational Analysis* (3rd printing), Berlin: Springer-Verlag.

[145] 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.

[146] Saad, Y. (2003) *Iterative Methods for Sparse Linear Systems* (2nd edition). Philadelphia: SIAM Publications.

[147] Schmidlin, P. (1972) "Iterative separation of sections in tomographic scintigrams." *Nuklearmedizin* **11**, pp. 1–16.

[148] Shepp, L., and Vardi, Y. (1982) "Maximum likelihood reconstruction for emission tomography." *IEEE Transactions on Medical Imaging*, **MI-1**, pp. 113–122.

[149] Stark, H., and Yang, Y. (1998) *Vector Space Projections. A Numerical Approach to Signal and Image processing, Neural Nets and Optics*, JNew York: John Wiley and Sons.

[150] Stark, H., and Woods, J. (2002) *Probability and Random Processes, with Applications to Signal Processing*. Upper Saddle River, NJ: Prentice-Hall.

[151] Tanabe, K. (1971) "Projection method for solving a singular system of linear equations and its applications." *Numer. Math.* **17**, pp. 203–214.

[152] Teboulle, M. (1992) "Entropic proximal mappings with applications to nonlinear programming." *Mathematics of Operations Research*, **17(3)**, pp. 670–690.

[153] 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.

[154] 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.

[155] Wright, M. (2009) "The dual flow between linear algebra and optimization." view-graphs of talk given at the History of Numerical Linear Algebra Minisymposium - Part II, SIAM Conference on Applied Linear Algebra, Monterey, CA, October 28, 2009.

[156] Wu, C.F.J. (1983) "On the convergence properties of the EM algorithm." *Annals of Stat.* **11**, pp. 95–103.

[157] Yang, Q. (2004) "The relaxed CQ algorithm solving the split feasibility problem." *Inverse Problems*, **20**, pp. 1261–1266.

[158] Yamada, I. (2001) "The hybrid steepest descent method for the variational inequality problem over the intersection of fixed points of nonexpansive mappings." in [30], pp. 473–504.

[159] Yamada, I., and Ogura, N. (2004) "Hybrid steepest descent method for the variational inequality problem over the fixed point set of certain quasi-nonexpansive mappings." Optimiz., **25**, pp. 619–655.

Index