Iterative Convex Optimization Algorithms; Part Two: Without the Baillon–Haddad Theorem

Charles L. Byrne*

February 24, 2015

Abstract

Let $C \subseteq X$ be a nonempty subset of an arbitrary set X and $f : X \to \mathbb{R}$. The problem is to minimize f over C. In auxiliary-function (AF) minimization we minimize $G_k(x) = f(x) + g_k(x)$ over x in C to get x^k, where $g_k(x) \geq 0$ for all x and $g_k(x^{k-1}) = 0$. Then the sequence $\{f(x^k)\}$ is nonincreasing and converges to some $\beta^* \geq -\infty$. A wide variety of iterative optimization methods are either in the AF class or can be reformulated to be in that class, including forward-backward splitting, barrier-function and penalty-function methods, alternating minimization, majorization minimization (MM) (or optimality transfer), cross-entropy minimization, and proximal minimization algorithms (PMA). The MM class and the PMA class are equivalent.

In order to have the sequence $\{f(x^k)\}$ converge to β, the infimum of $f(x)$ over x in C, we need to impose additional restrictions. An AF algorithm is said to be in the SUMMA class if, for all x, we have the SUMMA Inequality: $G_k(x) - G_k(x^k) \geq g_{k+1}(x)$. Then $\{f(x^k)\} \downarrow \beta$. Proximal minimization algorithms using Bregman distances (PMAB) form a large subclass of SUMMA. Not all PMA are PMAB, and so not all PMA are SUMMA. The PMA discussed by Auslender and Teboulle does have $\beta^* = \beta$, but appears not to be in the SUMMA class.

Here we generalize the SUMMA Inequality to obtain a wider class of algorithms that also contains the PMA of Auslender and Teboulle. Algorithms are said to be in the SUMMA2 class if, for each sequence $\{x^k\}$ of iterates, there are functions $h_k : X \to \mathbb{R}_+$ with

$$h_k(x) - h_{k+1}(x) \geq f(x^k) - f(x),$$

for all x and k. The algorithms of Auslender and Teboulle are in SUMMA2 and for all algorithms in SUMMA2 we have $\{f(x^k)\} \downarrow \beta$.

*Charles_Byrne@uml.edu, Department of Mathematical Sciences, University of Massachusetts Lowell, Lowell, MA 01854

1 Background

I began studying iterative algorithms around 1990, working with colleagues in the Department of Nuclear Medicine, University of Massachusetts Medical School. They introduced me to the EM algorithm, which they were starting to use for SPECT image reconstruction. Eventually, I noticed that the EM algorithm is closely related to the simultaneous MART (SMART) algorithm, and both are best studied using the Kullback–Leibler, or cross-entropy, distance measure.

1.1 The Kullback–Leibler Distance

For $a > 0$ and $b > 0$, let

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

$KL(a, 0) = +\infty$, and $KL(0, b) = b$. Then extend the KL distance component-wise to nonnegative vectors x and z, obtaining

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

Then $KL(x, z) \geq 0$ and $KL(x, z) = 0$ if and only if $x = z$.

1.2 The EM and SMART Algorithms

Let $y \in \mathbb{R}^I$ be a positive vector, and P an I by J matrix with nonnegative entries, whose columns sum to one. The EM algorithm minimizes the function $g(x) = KL(y, Px)$ over $x \geq 0$, while the SMART algorithm minimizes $f(x) = KL(Px, y)$. In [10] I developed the EM and SMART algorithms in tandem, to reveal both their similarities and subtle differences. The main tool that I used there was the *alternating minimization* (AM) method of Csiszár and Tusnády [28]. As the authors of [41] remark, the geometric argument in [28] is "deep, though hard to follow".

Shortly thereafter, I discovered that the iterative step of the SMART algorithm could be formulated as follows: for $k = 1, 2, ...$, minimize

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

to get

$$x_j^k = x_j^{k-1} \exp\Big(\sum_{i=1}^{I} P_{i,j} \log \frac{y_i}{(Px^{k-1})_i}\Big),$$

for $j = 1, ..., J$. This reminded me of the *sequential unconstrained minimization* (SUM) method of Fiacco and McCormick [29]. The SUM is used to minimize a

function $f : \mathcal{H} \to \mathbb{R}$ over a subset C. At each step of the SUM we minimize $f(x) + g_k(x)$ to get x^k. In SUM the $g_k(x)$ are selected to incorporate the constraint that x lie in C; barrier-function and penalty-function methods are the best known examples. In the case of SMART, the function

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

not only forces x to be nonnegative, but it also makes it possible to obtain the next iterate x^k in closed form.

1.3 The Landweber Algorithm

The Landweber (LW) algorithm minimizes the function

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

over all x in $\mathbb{R}^J$. The LW algorithm has the form of a gradient-descent method:

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

for appropriate $\gamma > 0$. The iterative step can be obtained by minimizing

$$f(x) + \frac{1}{2\gamma}\|x - x^{k-1}\|^2 - \frac{1}{2}\|Ax - Ax^{k-1}\|^2.$$

This is in the form of SUM, although there is no constraint to be imposed. The function

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

serves here only to provide a closed-form expression for the x^k.

1.4 The Projected Landweber Algorithm

The Projected Landweber (PLW) algorithm minimizes $f(x) = \frac{1}{2}\|Ax - b\|^2$ over x in C, where C is a nonempty, closed , convex subset of $\mathbb{R}^J$. We can obtain the iterative step of the PLW by minimizing

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

over x in C, to get

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

Note that the SUM approach is used here only to obtain a closed-form expression for x^k, not to impose the constraint.

1.5 What This Suggested

These three examples of the use of the SUM formalism suggested to me that it would be helpful to consider iterative methods in which, at each step, we minimize

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

to get x^k. If there are constraints, the $g_k(x)$ need not involve the constraints, and the $G_k(x)$ would still be minimized over C. The functions $g_k(x)$ would be selected to control the behavior of the sequences $\{f(x^k)\}$ and $\{x^k\}$.

The convergence of both the LW and PLW can be established using the Baillon–Haddad Theorem and the Krasnosel'skii–Mann Theorem for averaged operators. In the absence of an elementary proof of the BH Theorem, however, there was some advantage to exploring the SUM approach.

2 Overview

The Baillon–Haddad (BH) Theorem asserts that, if the gradient operator of a differentiable convex function on a Hilbert space is nonexpansive, then it is firmly nonexpansive. This theorem provides an important link between iterative fixed-point algorithms and convex optimization. Until last year [18, 19, 20], all the published proofs of the BH Theorem were non-elementary. For this reason, we investigate proofs of convergence of certain iterative optimization algorithms that do not rely on the BH Theorem and the Krasnosel'skii–Mann Theorem for averaged operators.

The basic problem we consider in this note 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 this note.

We begin by describing *auxiliary-function* (AF) methods for iterative optimization, and then focus on those AF algorithms that are in the SUMMA class. When the sequence $\{x^k\}$ is generated by an algorithm in the SUMMA class, we know that $\{f(x^k)\}$ converges to $\beta = \inf_{x \in C} f(x)$. As we shall see, a wide variety of iterative algorithms are in the SUMMA class, including proximal minimization algorithms

using Bregman distances (PMAB). A particular case of the PMAB is the *forward-backward splitting* (FBS) method. Convergence of the FBS algorithm is shown using the formalism of the PMAB, and without using the BH Theorem.

3 Auxiliary-Function Methods

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

$$G_k(x) = f(x) + g_k(x) \tag{3.1}$$

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 3.1 *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 in the next section.

3.1 Barrier- and Penalty-Function Algorithms

Perhaps the best known examples of AF methods are the *sequential unconstrained minimization* (SUM) methods discussed by Fiacco and McCormick in their classic book [29]. 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 [29] barrier-function methods are called *interior-point methods*, while penalty-function methods are called *exterior-point methods*.

In both the barrier- and penalty-function methods the auxiliary functions involve the subset C. In AF methods generally, however, this need not be the case, As we shall see, the auxiliary functions can sometimes be unrelated to C and selected to provide a closed-form expression for x^k.

3.2 Proximal Minimization Algorithms

The proximal minimization algorithms (PMA) form a large subclass of AF methods. In PMA we minimize $f(x)+d(x,x^{k-1})$ to get x^k, where $d(x,y) \geq 0$, and $d(x,x) = 0$ for all x. The *majorization minimization* (MM) method in statistics [33, 26], also called *optimization transfer*, is not typically formulated as an AF method, but it is one. The MM method is the following. Assume that there is a function $g(x|y) \geq f(x)$, for all x and y, with $g(y|y) = f(y)$. Then, for each k, minimize $g(x|x^{k-1})$ to get x^k. The MM methods and the PMA methods are equivalent; given $g(x|y)$, define $d(x,y) \doteq g(x|y) - f(x)$ and given $d(x,y)$, define $g(x|y) \doteq f(x) + d(x,y)$.

4 The SUMMA Class

An AF method is said to be in the SUMMA class if the SUMMA Inequality holds:

$$G_k(x) - G_k(x^k) \geq g_{k+1}(x), \tag{4.1}$$

for all $x \in X$. 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, including barrier-function methods, fall into the SUMMA class [15, 19].

We have the following theorem.

Theorem 4.1 *For $k = 1, 2, \ldots$ let x^k minimize $G_k(x) = f(x) + g_k(x)$ over $x \in C$, where the $g_k(x)$ satisfy the SUMMA inequality (4.1). Then the sequence $\{f(x^k)\}$ converges to $\beta = \inf_{x \in C} f(x)$.*

Proof: We know that the sequence $\{f(x^k)\}$ is decreasing, and therefore must converge to some β^*. Suppose that $\beta^* > \beta$. Then there is $z \in C$ such that

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

From

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

it follows that $\{g_k(z)\}$ is a decreasing sequence of nonnegative terms whose successive differences remain bounded away from zero, which is a contradiction. ∎

5 The PMAB Class

Proximal minimization algorithms using Bregman distances (PMAB) form an important subclass of the SUMMA class. We assume here that the function f is convex, but not necessarily differentiable. For general convex f we need the subdifferential of f at x, given by

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

and the fact that, if $x = z$ minimizes a convex function f, then $0 \in \partial f(z)$. A function $f : \mathbb{R}^N \to \mathbb{R}$ is convex if and only if $\partial f(x)$ is not empty, for all x.

5.1 The PMAB

Let $f : \mathbb{R}^N \to (-\infty, +\infty]$ be a convex function. Let h be another convex function, with effective domain $D = \{x | h(x) < +\infty\}$, 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 $C = \overline{D}$.

A function $f : \mathbb{R}^N \to (-\infty, +\infty]$ is *proper* if its essential domain $\text{dom}(f) = \{x | f(x) < +\infty\}$ is nonempty. If f is convex, then f is continuous on the interior of D. Therefore, if $f : \mathbb{R}^N \to \mathbb{R}$ is convex, it is continuous.

At the kth step of a PMAB algorithm [22, 23], we minimize the function

$$G_k(x) = f(x) + D_h(x, x^{k-1}), \tag{5.1}$$

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

$$g_k(x) = D_h(x, x^{k-1}) \tag{5.2}$$

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

$$G_k(x) - G_k(x^k) \geq D_h(x, x^k) = g_{k+1}(x) \geq 0, \tag{5.3}$$

so the PMAB is in the SUMMA class.

5.2 Some Computational Difficulties

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

$$\nabla h(x^{k-1}) - \nabla h(x^k) \in \partial f(x^k), \tag{5.4}$$

where the set

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

is the subdifferential of f at x. When $f(x)$ is differentiable, we must solve

$$\nabla f(x^k) + \nabla h(x^k) = \nabla h(x^{k-1}). \tag{5.5}$$

A modification of the PMAB, called the IPA for *interior-point algorithm* [11, 15], is designed to overcome these computational obstacles. We discuss the IPA later in this chapter. Another modification of the PMAB that is similar to the IPA is the *forward-backward splitting* (FBS) method to be discussed in a later section.

5.3 All PMAB are SUMMA

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

Lemma 5.1 *For each k we have*

$$G_k(x) - G_k(x^k) \geq D_h(x, x^k) = g_{k+1}(x). \tag{5.6}$$

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

$$0 \in \partial f(x^k) + \nabla h(x^k) - \nabla h(x^{k-1}), \tag{5.7}$$

so that

$$\nabla h(x^{k-1}) = u^k + \nabla h(x^k), \tag{5.8}$$

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 (5.8), to get

$$G_k(x) - G_k(x^k) = f(x) - f(x^k) - \langle u^k, x - x^k \rangle + D_h(x, x^k). \tag{5.9}$$

Therefore,

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

since u^k is in $\partial f(x^k)$. ∎

Convergence of the PMAB can be established, with some restrictions placed on the functions f and h (see [19]).

5.4 The IPA

The IPA is a modification of the PMAB designed to overcome some of the computational obstacles encountered in the PMAB [11, 15]. At the kth step of the PMAB we must solve the equation

$$\nabla f(x^k) + \nabla h(x^k) = \nabla h(x^{k-1}) \tag{5.10}$$

for x^k, where, for notational convenience, we assume that both f and h are differentiable. Solving Equation (5.10) 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 (5.10) now reads

$$\nabla a(x^k) = \nabla a(x^{k-1}) - \nabla f(x^{k-1}), \tag{5.11}$$

and we choose $a(x)$ so that Equation (5.11) is easily solved. We turn now to several examples of the IPA.

5.5 Projected Gradient Descent

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

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2. \tag{5.12}$$

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

$$a(x) = \frac{1}{2\gamma}\|x\|_2^2; \tag{5.13}$$

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

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

$$f(x) + \frac{1}{2\gamma}\|x - x^{k-1}\|_2^2 - D_f(x, x^{k-1}), \tag{5.14}$$

over $x \in C$.

For x^k to be the minimizer of $G_k(x)$ over $x \in C$, x^k must satisfy the inequalities

$$\langle \nabla G_k(x^k), c - x^k \rangle \geq 0,$$

for all c in C. Therefore, the solution x^k is in C and satisfies the inequality

$$\langle x^k - (x^{k-1} - \gamma \nabla f(x^{k-1})), c - x^k \rangle \geq 0, \tag{5.15}$$

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

$$x^k = P_C(x^{k-1} - \gamma \nabla f(x^{k-1})); \tag{5.16}$$

here P_C denotes the orthogonal projection onto C. This is the projected gradient descent algorithm. We obtain convergence of the projected gradient descent algorithm as a particular case of the forward-backward splitting method, to be proved in a later section. Note that the auxiliary function

$$g_k(x) = \frac{1}{2\gamma}\|x - x^{k-1}\|_2^2 - D_f(x, x^{k-1}) \tag{5.17}$$

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

$$x^k = x^{k-1} - \gamma \nabla f(x^{k-1}); \tag{5.18}$$

this is the gradient descent algorithm.

5.6 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

$$f(x) = \frac{1}{2}\|Ax - b\|_2^2, \tag{5.19}$$

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

$$\nabla f(x) = A^T(Ax - b) \tag{5.20}$$

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

$$D_f(x, z) = \frac{1}{2}\|Ax - Az\|_2^2. \tag{5.21}$$

We let

$$a(x) = \frac{1}{2\gamma}\|x\|_2^2, \tag{5.22}$$

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

$$G_k(x) = f(x) + D_h(x, x^{k-1}) \tag{5.23}$$

over $x \in C$ to get

$$x^k = P_C(x^{k-1} - \gamma A^T(Ax^{k-1} - b)); \tag{5.24}$$

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

6 Forward-Backward Splitting Algorithms

The *forward-backward splitting* (FBS) methods [27, 17] form a broad class of SUMMA algorithms closely related the IPA. Note that minimizing $G_k(x)$ in Equation (5.1) over $x \in C$ is equivalent to minimizing

$$G_k(x) = \iota_C(x) + f(x) + D_h(x, x^{k-1}) \tag{6.1}$$

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

6.1 The FBS Algorithm

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

$$\|\nabla f_2(x) - \nabla f_2(y)\|_2 \leq L\|x - y\|_2, \tag{6.2}$$

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

$$G_k(x) = f_1(x) + f_2(x) + \frac{1}{2\gamma}\|x - x^{k-1}\|_2^2 - D_{f_2}(x, x^{k-1}), \tag{6.3}$$

over all $x \in \mathbb{R}^N$, where $0 < \gamma < \frac{1}{2\gamma}$.

6.2 Moreau's Proximity Operators

Following Combettes and Wajs [27], we say that the *Moreau envelope* of index $\gamma > 0$ of the convex function $f : \mathbb{R}^N \to \mathbb{R}$ is the convex function

$$g(x) = \inf\{f(y) + \frac{1}{2\gamma}||x - y||_2^2\}, \tag{6.4}$$

with the infimum taken over all y in $\mathbb{R}^N$ [36, 37, 38]. In Rockafellar's book [39] and elsewhere, it is shown that the infimum is attained at a unique y, usually denoted $\text{prox}_{\gamma f}(x)$. The proximity operators $\text{prox}_{\gamma f}(\cdot)$ are firmly nonexpansive [27]; 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 [8]. 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 $z - x \in \partial f(x)$.

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 [27].

Let $f : \mathbb{R}^N \to \mathbb{R}$ 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

$$x^k = \text{prox}_{\gamma f_1}\Big(x^{k-1} - \gamma \nabla f_2(x^{k-1})\Big). \tag{6.5}$$

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 $(0, 1/L]$.

6.3 Convergence of the FBS algorithm

Let $f : \mathbb{R}^N \to \mathbb{R}$ 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 (6.5) and let $0 < \gamma \leq 1/L$.

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

$$G_k(x) = f(x) + \frac{1}{2\gamma}\|x - x^{k-1}\|_2^2 - D_{f_2}(x, x^{k-1}), \tag{6.6}$$

where

$$D_{f_2}(x, x^{k-1}) = f_2(x) - f_2(x^{k-1}) - \langle \nabla f_2(x^{k-1}), x - x^{k-1} \rangle. \tag{6.7}$$

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

The auxiliary function

$$g_k(x) = \frac{1}{2\gamma}\|x - x^{k-1}\|_2^2 - D_{f_2}(x, x^{k-1}) \tag{6.8}$$

can be rewritten as

$$g_k(x) = D_h(x, x^{k-1}), \tag{6.9}$$

where

$$h(x) = \frac{1}{2\gamma}\|x\|_2^2 - f_2(x). \tag{6.10}$$

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

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

$$\langle \nabla h(x) - \nabla h(y), x - y \rangle \geq 0, \tag{6.11}$$

for all x and y. This is equivalent to

$$\frac{1}{\gamma}\|x - y\|_2^2 - \langle \nabla f_2(x) - \nabla f_2(y), x - y \rangle \geq 0. \tag{6.12}$$

Since ∇f_2 is L-Lipschitz, the inequality (6.12) holds for $0 < \gamma \leq 1/L$.

Lemma 6.1 *The x^k that minimizes $G_k(x)$ over x is given by Equation (6.5).*

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

Proof: A relatively simple calculation shows that

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

$$\Big(f_1(x) - f_1(x^k) - \frac{1}{\gamma}\langle (x^{k-1} - \gamma \nabla f_2(x^{k-1})) - x^k, x - x^k \rangle\Big). \tag{6.13}$$

Since

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

it follows that

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

Therefore,

$$G_k(x) - G_k(x^k) \geq \frac{1}{2\gamma} \|x - x^k\|_2^2 \geq g_{k+1}(x). \tag{6.14}$$

Therefore, the inequality in (4.1) 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

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

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

7 Some Examples of the FBS Algorithms

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

7.1 Projected Gradient Descent

Let C be a nonempty, closed convex subset of $\mathbb{R}^N$ 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 (6.5) becomes

$$x^k = P_C\Big(x^{k-1} - \gamma \nabla f_2(x^{k-1})\Big). \tag{7.1}$$

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

7.2 The CQ Algorithm

Let A be a real I by J matrix, $C \subseteq \mathbb{R}^N$, 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

$$f_2(x) = \frac{1}{2}\|P_Q Ax - Ax\|_2^2 \tag{7.2}$$

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

$$\nabla f_2(x) = A^T(I - P_Q)Ax. \tag{7.3}$$

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

$$x^k = P_C\Big(x^{k-1} - \gamma A^T(I - P_Q)Ax^{k-1}\Big); \tag{7.4}$$

this iterative method was called the CQ-algorithm in [12, 13]. The sequence $\{x^k\}$ converges to a solution whenever f_2 has a minimum on the set C, for $0 < \gamma \leq 1/L$.

In [25, 24] the CQ algorithm was extended to a multiple-sets algorithm and applied to the design of protocols for intensity-modulated radiation therapy.

7.3 The Projected Landweber Algorithm

The problem is to minimize the function

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

over $x \in C$. This is a special case of the SFP and we can use the CQ-algorithm, with $Q = \{b\}$. The resulting iteration is the projected Landweber algorithm [7]; when $C = \mathbb{R}^N$ it becomes the Landweber algorithm [32].

8 Alternating Minimization

The main reference for alternating minimization is the paper [28] of Csiszár and Tusnády. As the authors of [41] remark, the geometric argument in [28] 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 [28] holds fall into the SUMMA class (see [16]).

8.1 The AM Framework

Suppose that P and Q are arbitrary non-empty sets and the function $\Theta(p, q)$ satisfies $-\infty < \Theta(p, q) \leq +\infty$, for each $p \in P$ and $q \in Q$. We assume that, for each $p \in P$, there is $q \in Q$ with $\Theta(p, q) < +\infty$. Therefore, $\beta = \inf_{p \in P,\, q \in Q} \Theta(p, q) < +\infty$. 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 $\Theta(p^n, q^n) \downarrow \beta$.

8.2 The AM Iteration

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

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

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

$$\Theta(p^n, q^n) \geq \Theta(p^{n+1}, q^n) \geq \Theta(p^{n+1}, q^{n+1}). \tag{8.1}$$

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

8.3 The Five-Point Property for AM

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

The Five-Point Property

$$\Theta(p, q) + \Theta(p, q^{n-1}) \geq \Theta(p, q^n) + \Theta(p^n, q^{n-1}). \tag{8.2}$$

9 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 [28].

Theorem 9.1 *If the five-point property holds then* $\beta^* = \beta$.

9.1 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

$$G_n(p) = \Theta(p, q^{n-1}) = \Theta(p, q(p)) + \Big(\Theta(p, q^{n-1}) - \Theta(p, q(p))\Big) \tag{9.3}$$

to get p^n. With

$$g_n(p) = \Big(\Theta(p, q^{n-1}) - \Theta(p, q(p))\Big) \geq 0, \tag{9.4}$$

we can write

$$G_n(p) = f(p) + g_n(p). \tag{9.5}$$

According to the five-point property, we have

$$G_n(p) - G_n(p^n) \geq \Theta(p, q^n) - \Theta(p, q(p)) = g_{n+1}(p). \tag{9.6}$$

It follows that AM is a member of the SUMMA class, whenever the five-point property holds, in which case the sequence $\{\Theta(p^n, q^n)\}$ converges to β.

10 The SUMMA2 Class

A consequence of the SUMMA Inequality is that, for all x in C,

$$g_k(x) - g_{k+1}(x) \geq f(x^k) - f(x). \tag{10.1}$$

This inequality is central to the proof that $\beta^* = \beta$ for algorithms in the SUMMA class. However, not all PMA methods with $\beta^* = \beta$ are in the SUMMA class; those discussed by Auslender and Teboulle in [1] appear not to be in SUMMA. We seek to widen the SUMMA class to include the Auslender-Teboulle methods.

10.1 The SUMMA2 Inequality

An AF algorithm is said to be in the SUMMA2 class if, for each sequence $\{x^k\}$ generated by the algorithm, each x^k is in C and there are functions $h_k : X \to \mathbb{R}_+$ such that, for all $x \in C$, we have

$$h_k(x) - h_{k+1}(x) \geq f(x^k) - f(x). \tag{10.2}$$

Any algorithm in the SUMMA class is in the SUMMA2 class; use $h_k = g_k$. In addition, as we shall show, the proximal minimization method of Auslender and Teboulle [1] is also in the SUMMA2 class. As in the SUMMA case, we must have $\beta^* = \beta$, since otherwise the successive differences of the sequence $\{h_k(z)\}$ would be bounded below by $\beta^* - f(z) > 0$. It is helpful to note that the functions h_k need not be the g_k, and we do not require that $h_k(x^{k-1}) = 0$.

10.2 The Auslender-Teboulle Class

In [1] Auslender and Teboulle take C to be a closed, nonempty, convex subset of $\mathbb{R}^N$, with interior U. At the kth step of their method one minimizes a function

$$G_k(x) = f(x) + d(x, x^{k-1}) \tag{10.3}$$

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

$$\langle \nabla_1 d(b, a), c - b \rangle \leq H(c, a) - H(c, b), \tag{10.4}$$

for all c in U.

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

$$\langle \nabla_1 d(b,a), c-b \rangle = D_h(c,a) - D_h(c,b) - D_h(b,a) \tag{10.5}$$

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 [1]. The method of Auslender and Teboulle seems not to be a particular case of SUMMA. However, it is in the SUMMA2 class, as we now show.

Since x^k minimizes $f(x) + d(x, x^{k-1})$, it follows that

$$0 \in \partial f(x^k) + \nabla_1 d(x^k, x^{k-1}),$$

so that

$$-\nabla_1 d(x^k, x^{k-1}) \in \partial f(x^k).$$

We then have

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

Using the associated induced proximal distance H, we obtain

$$f(x^k) - f(x) \leq H(x, x^{k-1}) - H(x, x^k).$$

Therefore, this method is in the SUMMA2 class, with the choice of $h_k(x) = H(x, x^{k-1})$. Consequently, we have $\beta^* = \beta$ for these algorithms.

References

1. Auslender, A., and Teboulle, M. (2006) "Interior gradient and proximal methods for convex and conic optimization." *SIAM Journal on Optimization*, **16(3)**, pp. 697–725.

2. 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.

3. 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.

4. Bauschke, H., and Borwein, J. (1996) "On projection algorithms for solving convex feasibility problems." *SIAM Review*, **38 (3)**, pp. 367–426.

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

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

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

8. 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.

9. 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.

10. 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.

11. Byrne, C. (2001) "Bregman-Legendre multi-distance projection algorithms for convex feasibility and optimization." in [9], pp. 87–100.

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

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

14. Byrne, C. (2007) *Applied Iterative Methods*. Wellesley, MA: A K Peters.

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

16. 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.

17. 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.

18. Byrne, C. (2014) "On a generalized Baillon–Haddad Theorem for convex functions on Hilbert space." to appear in the *Journal of Convex Analysis*.

19. Byrne, C. (2014) *Iterative Optimization in Inverse Problems*. Boca Raton, FL: CRC Press.

20. Byrne, C. (2014) *A First Course in Optimization*. Boca Raton, FL: CRC Press.

21. Byrne, C. (2015) "Auxiliary-function methods in iterative optimization." submitted for publication.

22. Censor, Y., and Zenios, S.A. (1992) "Proximal minimization algorithm with D-functions." *Journal of Optimization Theory and Applications*, **73(3)**, pp. 451–464.

23. Censor, Y. and Zenios, S.A. (1997) *Parallel Optimization: Theory, Algorithms and Applications*. New York: Oxford University Press.

24. 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.

25. 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.

26. Chi, E., Zhou, H., and Lange, K. (2014) "Distance Majorization and Its Applications." Mathematical Programming, **146 (1-2)**, pp. 409–436.

27. Combettes, P., and Wajs, V. (2005) "Signal recovery by proximal forward-backward splitting." *Multiscale Modeling and Simulation*, **4(4)**, pp. 1168–1200.

28. Csiszár, I. and Tusnády, G. (1984) "Information geometry and alternating minimization procedures." *Statistics and Decisions* **Supp. 1**, pp. 205–237.

29. Fiacco, A., and McCormick, G. (1990) *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*. Philadelphia, PA: SIAM Classics in Mathematics (reissue).

30. Golshtein, E., and Tretyakov, N. (1996) *Modified Lagrangians and Monotone Maps in Optimization*. New York: John Wiley and Sons.

31. Krasnosel'skii, M. (1955) "Two observations on the method of sequential approximations." *Uspeki Mathematicheskoi Nauki* (in Russian), **10(1)**.

32. Landweber, L. (1951) "An iterative formula for Fredholm integral equations of the first kind." *Amer. J. of Math.* **73**, pp. 615–624.

33. Lange, K., Hunter, D., and Yang, I. (2000) "Optimization transfer using surrogate objective functions (with discussion)." *J. Comput. Graph. Statist.*, **9**, pp. 1–20.

34. Lucet, Y. (2010) "What shape is your conjugate? A survey of computational convex analysis and its applications." *SIAM Review*, **52(3)**, pp. 505–542.

35. Mann, W. (1953) "Mean value methods in iteration." *Proc. Amer. Math. Soc.* **4**, pp. 506–510.

36. 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.

37. Moreau, J.-J. (1963) "Propriétés des applications 'prox'." *C.R. Acad. Sci. Paris Sér. A Math.*, **256**, pp. 1069–1071.

38. Moreau, J.-J. (1965) "Proximité et dualité dans un espace hilbertien." *Bull. Soc. Math. France*, **93**, pp. 273–299.

39. Rockafellar, R. (1970) *Convex Analysis*. Princeton, NJ: Princeton University Press.

40. Rockafellar, R.T. and Wets, R. J-B. (2009) *Variational Analysis* (3rd printing), Berlin: Springer-Verlag.

41. 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.