Auxiliary-Function Methods in Iterative Optimization

Charles L. Byrne*

April 6, 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 X 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. 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 (optimality transfer), cross-entropy minimization, and proximal minimization methods. 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 $\{f(x^k)\} \downarrow \beta$. Here we generalize the SUMMA Inequality to obtain a wider class of algorithms that also contains the proximal minimization methods of Auslender and Teboulle. Algorithms are said to be in the SUMMA2 class if there are functions $h_k : X \to \mathbb{R}_+$ such that $h_k(x) - h_{k+1}(x) \geq f(x^k) - f(x)$ for all x in C. Once again, we have $\{f(x^k)\} \downarrow \beta$.

Key Words: Sequential unconstrained minimization; forward-backward splitting; proximal minimization; Bregman distances.

2000 Mathematics Subject Classification: Primary 47H09, 90C25; Secondary 26A51, 26B25.

1 Auxiliary-Function Methods

The basic problem we consider in this paper 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,

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

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 to be discussed here.

The algorithms we consider are of the sequential minimization type. 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 $x^k \in C$. If C is a proper subset of X we replace $f(x)$ with $f(x) + \iota_C(x)$, where $\iota_C(x) = 0$, for $x \in C$, and $\iota_C(x) = +\infty$, otherwise; then the minimization is automatically over $x \in C$. In some cases, but not always, the functions $g_k(x)$ may be used to incorporate the constraint that $f(x)$ is to be minimized over $x \in C$. As we shall see, 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 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 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 and converges to some $\beta^* \geq -\infty$.*

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 additional restrictions.

Perhaps the best known examples of AF methods are the *sequential unconstrained minimization* (SUM) methods discussed by Fiacco and McCormick in their classic book [20]. They focus on barrier-function and penalty-function algorithms, which are not usually presented in AF form, but can be reformulated as members of the

AF class. In [20] barrier-function methods are called *interior-point methods*, while penalty-function methods are called *exterior-point methods*. 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 (optimality transfer), cross-entropy minimization, and proximal minimization 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.

2 The SUMMA Class

Simply asking that the sequence $\{f(x^k)\}$ be nonincreasing is usually not enough. We want $\{f(x^k)\} \downarrow \beta = \inf_{x \in C} f(x)$. This occurs in most of the examples mentioned above. In [9] it was shown that, if the auxiliary functions g_k are selected so as to satisfy the SUMMA Inequality,

$$G_k(x) - G_k(x^k) \geq g_{k+1}(x), \tag{2.1}$$

for all $x \in C$, then $\beta^* = \beta$. Although there are many iterative algorithms that satisfy the SUMMA Inequality, and are therefore in the SUMMA class, some important methods that are not in this class still have $\beta^* = \beta$; one example is the proximal minimization method of Auslender and Teboulle [1]. This suggests that the SUMMA class, large as it is, is still unnecessarily restrictive.

One consequence of the SUMMA Inequality is

$$g_k(x) - g_{k+1}(x) \geq f(x^k) - f(x), \tag{2.2}$$

for all $x \in C$. It follows from this that $\beta^* = \beta$. If this were not the case, then there would be $z \in C$ with

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

for all k. The sequence $\{g_k(z)\}$ would then be a nonincreasing sequence of nonnegative terms with the sequence of its successive differences bounded below by $\beta^* - f(z) > 0$. In order to widen the SUMMA class to include the proximal minimization method of Auslender and Teboulle we focus on generalizing the inequality (2.2).

3 The SUMMA2 Class

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

$$h_k(x) - h_{k+1}(x) \geq f(x^k) - f(x). \tag{3.1}$$

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

4 Proximal Minimization Algorithms

Let $d : X \times X \to \mathbb{R}_+$ be a "distance", meaning simply that $d(x, y) = 0$ if and only if $x = y$. An iterative algorithm is a *proximal minimization algorithm* (PMA) if, for each k, we minimize

$$G_k(x) = f(x) + d(x, x^{k-1}) \tag{4.1}$$

to get x^k. Clearly, any method in the PMA class is also an AF method.

4.1 Majorization Minimization

The *majorization minimization* (MM) method in statistics [23, 17], 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 $g(x|y) \doteq f(x) + d(x, y)$.

4.2 PMA with Bregman Distances

Let $\mathcal{H}$ be a Hilbert space, and $h : \mathcal{H} \to \mathbb{R}$ strictly convex and Gâteaux differentiable. The *Bregman distance* associated with h is

$$D_h(x, y) = h(x) - h(y) - \langle \nabla h(y), x - y \rangle. \tag{4.2}$$

Proximal minimization with Bregman distances (PMAB) applies to the minimization of a convex function $f : \mathcal{H} \to \mathbb{R}$. In [13, 14] Censor and Zenios discuss in detail the PMAB methods, which they call proximal minimization with D-functions.

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

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

where

$$\partial f(x) = \{u | f(y) - f(x) - \langle \nabla u, y - x \rangle \geq 0, \text{for all } y\}$$

is the subdifferential of f at x. In [9] it was shown that for the PMAB methods we have $u^k \in \partial f(x^k)$ such that

$$G_k(x) - G_k(x^k) = f(x) - f(x^k) - \langle u^k, x - x^k \rangle + D_h(x, x^k) \geq g_{k+1}(x), \tag{4.3}$$

for all x. Consequently, the SUMMA Inequality holds and all PMAB algorithms are in the SUMMA class.

4.3 The Forward-Backward Splitting Methods

The *forward-backward splitting* (FBS) methods discussed by Combettes and Wajs [18] form a particular subclass of the PMAB methods. The problem now is to minimize the function $f(x) = f_1(x) + f_2(x)$, where both $f_1 : \mathcal{H} \to (-\infty, +\infty]$ and $f_2 : \mathcal{H} \to (-\infty, +\infty]$ are lower semicontinuous, proper and convex, and f_2 is Gâteaux differentiable, with L-Lipschitz continuous gradient. Before we describe the FBS algorithm we need to recall Moreau's proximity operators.

Following Combettes and Wajs [18], 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

$$\text{env}_{\gamma f}(x) = \inf_{y \in \mathcal{H}} \{f(y) + \frac{1}{2\gamma} ||x - y||^2\}; \tag{4.4}$$

see also Moreau [24, 25, 26]. In Rockafellar's book [27] and elsewhere, it is shown that the infimum is attained at a unique y, usually denoted $\text{prox}_{\gamma f}(x)$. Proximity operators 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 $z - x \in \partial f(x)$.

In [18] the authors show, using the characterization of $\text{prox}_{\gamma f}$ given above, that x is a solution of this minimization problem if and only if

$$x = \text{prox}_{\gamma f_1}(x - \gamma \nabla f_2(x)). \tag{4.5}$$

This suggests to them the following FBS iterative scheme:

$$x^k = \text{prox}_{\gamma f_1}(x^{k-1} - \gamma \nabla f_2(x^{k-1})). \tag{4.6}$$

Basic properties and convergence of the FBS algorithm are then developed in [18].

In [11] we presented a simplified proof of convergence for the FBS algorithm. The basic idea used there is to formulate the FBS algorithm as a member of the PMAB class. An easy calculation shows that, if we minimize

$$G_k(x) = f_1(x) + f_2(x) + \frac{1}{2\gamma}\|x - x^{k-1}\|^2 - D_{f_2}(x, x^{k-1}), \tag{4.7}$$

we get x^k as described in Equation (4.6). The function

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

is convex and Gâteaux differentiable, when $0 < \gamma \leq \frac{1}{L}$, and

$$D_h(x, x^{k-1}) = \frac{1}{2\gamma}\|x - x^{k-1}\|^2 - D_{f_2}(x, x^{k-1}).$$

Therefore, the FBS method is in the PMAB class. A number of well known iterative algorithms are particular cases of the FBS.

4.4 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 (4.6) becomes

$$x^k = P_C\left(x^{k-1} - \gamma \nabla f_2(x^{k-1})\right). \tag{4.8}$$

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

4.5 The CQ Algorithm and Split Feasibility

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

$$f_2(x) = \frac{1}{2}\|P_Q Ax - Ax\|^2 \tag{4.9}$$

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

$$\nabla f_2(x) = A^T(I - P_Q)Ax. \tag{4.10}$$

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

$$x^k = P_C\left(x^{k-1} - \gamma A^T(I - P_Q)Ax^{k-1}\right); \tag{4.11}$$

this iterative method was called the CQ-algorithm in [7, 8]. The sequence $\{x^k\}$ converges to a solution whenever f_2 has a minimum on the set C, for $0 < \gamma \leq 1/L$.

If $Q = \{b\}$, then the CQ algorithm becomes the *projected Landweber* algorithm [3]. If, in addition, $C = \mathbb{R}^J$, then we get the Landweber algorithm [22]. In [15, 16] Yair Censor and his colleagues modified the CQ algorithm and applied it to derive protocols for intensity-modulated radiation therapy.

4.6 The PMA of Auslender and Teboulle

In [1] Auslender and Teboulle 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

$$G_k(x) = f(x) + d(x, x^{k-1}) \tag{4.12}$$

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

$$\langle \nabla_1 d(b, a), c - b\rangle \leq H(c, a) - H(c, b), \tag{4.13}$$

for all c in U.

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

$$\langle \nabla_1 d(b,a), c-b \rangle = D_h(c,a) - D_h(c,b) - D_h(b,a) \tag{4.14}$$

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.

It is interesting to note that the Auslender-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 $f(x)$ be convex; it does not even require any structure on X. The SUMMA2 approach is general enough to include both classes.

In the next few sections we consider several other optimization problems and iterative methods that are particular cases of the SUMMA class.

5 Barrier-Function and Penalty-Function Methods

Barrier-function methods and penalty-function methods for constrained optimization are not typically presented as AF methods [20]. However, barrier-function methods

can be reformulated as AF algorithms and shown to be members of the SUMMA class. Penalty-function methods can be rewritten in the form of barrier-function methods, permitting several facts about penalty-function algorithms to be obtained from related results on barrier-function methods.

5.1 Barrier-Function Methods

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 5.1 *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 5.1 depended heavily on the details of the barrier-function method. Now we reformulate the barrier-function method as an AF method.

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. A simple calculation shows that

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

for all $x \in X$. Therefore, barrier-function methods are particular cases of the SUMMA class.

5.2 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*.

We suppose that $f(x) \geq \alpha > -\infty$, for all x. Replacing $f(x)$ with $f(x) - \alpha$ if necessary, we may assume that $f(x) \geq 0$, for all x. 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 5.1 it follows that the sequence $\{p(x^k)\}$ is nonincreasing 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 can be 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 $\gamma = \beta - 1$.

6 Cross-Entropy Methods

For $a > 0$ and $b > 0$, let the cross-entropy or Kullback-Leibler (KL) distance [21] from a to b be

$$KL(a, b) = a \log \frac{a}{b} + b - a, \tag{6.1}$$

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

$$KL(x, z) = \sum_{j=1}^{J} KL(x_j, z_j). \tag{6.2}$$

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.

The *simultaneous multiplicative algebraic reconstruction technique* (SMART) for minimizing $f(x) = KL(Px, y)$ over nonnegative $x \in \mathbb{R}^J$ has the iterative step

$$x^k = x^{k-1} \exp\left(\sum_{i=1}^{I} P_{i,j} \log \frac{y_i}{(Px^{k-1})_i}\right), \tag{6.3}$$

under the assumption that all columns of the matrix P sum to one. In [4, 5, 6] it was shown that x^k can be obtained by minimizing

$$G_k(x) = KL(Px, y) + KL(x, x^{k-1}) - KL(Px, Px^{k-1}). \tag{6.4}$$

We have

$$KL(x, z) - KL(Px, Pz) = D_h(x, z), \tag{6.5}$$

for

$$h(x) = \sum_{j=1}^{J} (x_j \log x_j - x_j) - KL(Px, y),$$

which is convex and Gâteaux differentiable. Therefore, the SMART algorithm is a particular case of PMAB. The SMART sequence $\{x^k\}$ converges to the nonnegative minimizer of $KL(Px, y)$ for which $KL(x, x^0)$ is minimized. If the entries of the starting vector x^0 are all one, then the sequence $\{x^k\}$ converges to the minimizer of $KL(Px, y)$ with maximum Shannon entropy [4].

7 Alternating Minimization

In [6] the SMART and the related EMML algorithm [29] were derived in tandem using the *alternating minimization* (AM) approach of Csiszár and Tusnády [19]. The AM approach is the following.

Let $\Theta : X \times Y \to (-\infty, +\infty]$, where X and Y are arbitrary nonempty sets. In the AM approach we minimize $\Theta(x, y^{k-1})$ over $x \in X$ to get x^k and then minimize $\Theta(x^k, y)$ over $y \in Y$ to get y^k. We want

$$\{\Theta(x^k, y^k)\} \downarrow \beta = \inf\{\Theta(x, y) | x \in X, y \in Y\}. \tag{7.1}$$

In [19] Csiszár and Tusnády show that, if the function Θ possesses what they call the *five-point property*,

$$\Theta(x, y) + \Theta(x, y^{k-1}) \geq \Theta(x, y^k) + \Theta(x^k, y^{k-1}), \tag{7.2}$$

for all x, y, and k, then Equation (7.1) 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 x, and the five-point property for AM is precisely the SUMMA Inequality [10]. For each x select $y(x)$ for which $\Theta(x, y(x)) \leq \Theta(x, y)$ for all $y \in Y$. Then let $f(x) = \Theta(x, y(x))$.

8 Applying Alternating Minimization

In [2] Bauschke, Combettes and Noll consider the following problem: minimize the function

$$\Theta(x, y) = \Lambda(x, y) = \phi(x) + \psi(y) + D_f(x, y), \tag{8.1}$$

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

$$\beta = \inf_{(x,y)} \Lambda(x, y) > -\infty, \tag{8.2}$$

and seek a sequence $\{(x^k, y^k)\}$ such that $\{\Lambda(x^k, y^k)\}$ 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 $\{\Lambda(x^k, y^k)\} \downarrow \beta$. In [12]we obtained this result by showing that $\Lambda(x, y)$ has the five-point property whenever D_f is jointly convex.

From our previous discussion of AM, we conclude that the sequence $\{\Lambda(x^n, y^n)\}$ converges to β; this is Corollary 4.3 of [2].

This suggests another class of proximal minimization methods for which $\beta^* = \beta$. Suppose that $D_f(x, y)$ is a jointly convex Bregman distance. For each $k = 1, 2, ...,$, we minimize

$$G_k(x) = f(x) + D_f(x^{k-1}, x) \tag{8.3}$$

to get x^k. Then using the result from [2], we may conclude that $\beta^* = \beta$.

9 Summary

We have considered the problem of minimizing $f : X \to \mathbb{R}$ over x in C, a nonempty subset of the arbitrary set X. For $k = 1, 2, ...$ we minimize $G_k(x) = f(x) + g_k(x)$ to get x^k. For a sequence $\{x^k\}$ generated by an AF algorithm the sequence $\{f(x^k)\}$ is nonincreasing and converges to some $\beta^* \geq -\infty$. In addition, for AF algorithms in the SUMMA class we have $\{f(x^k)\} \downarrow \beta = \inf_{x \in C} f(x)$; so $\beta^* = \beta$.

The SUMMA class of algorithms is quite large, but there are algorithms not in the SUMMA class for which $\beta^* = \beta$; the proximal minimization method of Auslender and Teboulle [1] is an example. The SUMMA Inequality is sufficient to guarantee that $\beta^* = \beta$, but it is clearly overly restrictive. We extend the SUMMA class to the SUMMA2 class by generalizing the SUMMA Inequality and show that the methods of [1] are members of the larger SUMMA2 class.

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

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

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

5. 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.

6. 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.

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

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

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

10. 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.

11. 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.

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

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

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

15. 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.

16. 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.

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

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

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

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

21. Kullback, S. and Leibler, R. (1951) "On information and sufficiency." *Annals of Mathematical Statistics* **22**, pp. 79–86.

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

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

24. 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.

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

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

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

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

29. 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.