Alternating Minimization as Sequential Unconstrained Minimization: A Survey*†

Charles L. Byrne

communicated by Marc Teboulle

July 17, 2012

Abstract

Sequential unconstrained minimization is a general iterative method for minimizing a function over a given set. At each step of the iteration we minimize the sum of the objective function and an auxiliary function. The aim is to select the auxiliary functions so that, at least, we get convergence in function value to the constrained minimum. The SUMMA is a broad class of these methods for which such convergence holds. Included in the SUMMA class are the barrier-function methods, entropic and other proximal minimization algorithms, the simultaneous multiplicative algebraic reconstruction technique, and, after some reformulation, penalty-function methods. The alternating minimization method of Csiszár and Tusnády also falls within the SUMMA class, whenever their five-point property holds. Therefore, the expectation maximization maximum likelihood algorithm for the Poisson case is also in the SUMMA class.

Key Words: optimization; sequential unconstrained optimization; alternating minimization.

AMS Classification: 65K10; 90C51.

Accepted for publication: Journal of Optimization Theory and Applications

1 Introduction

The alternating minimization (AM) method of Csiszár and Tusnády [1] is a framework for minimizing a function of two separately constrained variables. When their five-point property holds, the function values converge monotonically to the infimum of the function values over the constraint sets.

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

†I wish to thank Professor Heinz Bauschke for calling to my attention the article [17] and to the anonymous reviewers for helpful suggestions.

It was noticed by Rockmore and Macovski [2] that the image reconstruction problems that arise in medical tomography can be formulated as statistical parameter estimation problems. Following up on this idea, Shepp and Vardi [3] suggested the use of the EM algorithm, called here the EMML algorithm, for solving the reconstruction problem in emission tomography. In [4] Lange and Carson presented an EM-type iterative method for transmission tomographic image reconstruction, and pointed out a gap in the convergence proof given in [3] for the emission case. In [5], Vardi, Shepp and Kaufman repaired the earlier proof, relying on techniques due to Csiszár and Tusnády [1]. In [6] Lange, Bahn and Little improved the transmission and emission algorithms, by including regularization to reduce the effects of noise. The question of uniqueness of the solution in the inconsistent case was resolved in [7, 8].

What is usually called the simultaneous multiplicative algebraic reconstruction technique (SMART) was discovered independently in 1972, by Darroch and Ratcliff [9], working in statistics, and by Schmidlin [10] in medical imaging. The SMART is best derived using the AM formalism and provides an example of alternating minimization having the five-point property. Details concerning the SMART can be found in [7, 11], and in the references therein.

At each step of a sequential unconstrained minimization algorithm, one minimizes the sum of the objective function and an auxiliary function. The auxiliary functions can be chosen to enforce constraints on the vector variable, or simply to allow each iterate to be obtained in closed form. When the auxiliary functions are properly selected, the constraints are enforced and the iterated function values converge to the infimum of the values of the function over the constraint set. A standard reference for these methods is the 1967 book by Fiacco and McCormick [12]. The SUMMA [13] is a broad class of sequential unconstrained minimization algorithms that includes barrier-function methods, proximal minimization with Bregman functions [14, 15, 16], the SMART, and, after some reformulation, penalty-function methods. The choice of the auxiliary functions in SUMMA guarantees the convergence of the iterated function values to the infimum.

We show here that the AM procedure discussed in [17] has the five-point property whenever the Bregman distance involved is jointly convex, and that all AM methods with the five-point property are members of the SUMMA class.

2 Alternating Minimization

The alternating minimization (AM) approach of Csiszár and Tusnády [1] provides a useful framework for the derivation of iterative optimization algorithms. In this section we discuss their five-point property and convergence for their AM algorithm.

2.1 The AM Framework

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

2.2 The AM Iteration

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

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

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

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

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

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

We know two things:

$$\Theta(p^{n+1}, q^n) - \Theta(p^{n+1}, q^{n+1}) \geq 0, \tag{2}$$

and

$$\Theta(p^n, q^n) - \Theta(p^{n+1}, q^n) \geq 0. \tag{3}$$

The inequality in (3) can be strengthened to

$$\Theta(p, q^n) - \Theta(p^{n+1}, q^n) \geq 0. \tag{4}$$

We need to make these inequalities more precise.

2.3 The Five-Point Property for AM

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

The Five-Point Property

$$\Theta(p, q) + \Theta(p, q^{n-1}) \geq \Theta(p, q^n) + \Theta(p^n, q^{n-1}). \tag{5}$$

2.4 The Main Theorem for AM

We want to find sufficient conditions for the sequence $\{\Theta(p^n, q^n)\}$ to converge to b; that is, for $B = b$. The following is the main result of [1].

Theorem 2.1 *If the five-point property holds, then $B = b$.*

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

$$\Theta(p', q^{n-1}) - \Theta(p^n, q^{n-1}) \geq \Theta(p', q^n) - \Theta(p', q'), \tag{6}$$

so that

$$\Theta(p', q^{n-1}) - \Theta(p', q^n) \geq \Theta(p^n, q^{n-1}) - \Theta(p', q') \geq 0. \tag{7}$$

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 (7) must therefore converge to zero, which is a contradiction. We conclude that $B = b$ whenever the five-point property holds in AM. □

2.5 The Three- and Four-Point Properties

In [1] 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

$$\Theta(p, q^n) - \Theta(p^{n+1}, q^n) \geq \Delta(p, p^{n+1}), \tag{8}$$

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

The Four-Point Property

$$\Delta(p, p^{n+1}) + \Theta(p, q) \geq \Theta(p, q^{n+1}), \tag{9}$$

for all p and q.

Clearly 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

$$\Delta(p, \tilde{p}) := \Theta(p, q(\tilde{p})) - \Theta(p, q(p)), \tag{10}$$

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

Next, we discuss the SMART and EMML algorithms, two important instances of alternating minimization.

2.6 The SMART

We consider now the *simultaneous multiplicative algebraic reconstruction technique* (SMART) as an example of AM.

Let y have only positive entries y_i, and the matrix P have only non-negative entries P_{ij}, normalized so that $\sum_{i=1}^{I} P_{ij} = 1$, for all $j = 1, ..., J$. The SMART iteration begins with a positive vector $x^0 \in \mathbb{R}^J$. Having found the vector x^{n-1}, the next vector in the SMART sequence is x^n, with entries given by

$$x_j^n = x_j^{n-1} \exp\Big(\sum_{i=1}^{I} P_{ij} \log\Big(\frac{y_i}{(Px^{n-1})_i}\Big)\Big). \tag{11}$$

The sequence $\{x^n\}$ converges to the non-negative minimizer of the function $KL(Px, y)$ for which $KL(x, x^0)$ is minimized [7]; here KL denotes the Kullback-Leibler distance between non-negative vectors [18]:

$$KL(x, z) = \sum_{j=1}^{J} x_j \log \frac{x_j}{z_j} + z_j - x_j. \tag{12}$$

In [7] it was shown that the SMART iteration can be obtained through AM and that the five-point property holds.

2.7 The EMML Algorithm

The *expectation maximization maximum likelihood* (EMML) method we discuss here is actually a special case of a more general approach to likelihood maximization, usually called the EM algorithm [19]; the book by McLachnan and Krishnan [20] is a good source for the history of this more general algorithm.

The EMML, as a statistical parameter estimation technique, was not originally thought to be connected to any system of linear equations. In [7] it was shown that the EMML algorithm minimizes the function $f(x) = KL(y, Px)$, over non-negative vectors x; consequently, when the non-negative system of linear equations $Px = y$ has a non-negative solution, the EMML converges to such a solution.

2.8 The EMML Iteration

The EMML iteration begins with a positive vector $x^0 \in \mathbb{R}^J$. Having found the vector x^n, the next vector in the EMML sequence is x^{n+1}, with entries given by

$$x_j^{n+1} = x_j^n \sum_{i=1}^{I} P_{ij}\Big(\frac{y_i}{(Px^n)_i}\Big). \tag{13}$$

The sequence $\{x^n\}$ converges to a non-negative minimizer of the function $KL(y, Px)$. In [5] it was shown that the EMML algorithm is a special case of AM and that the five-point property holds.

2.9 Alternating Bregman Distance Minimization

The general problem of minimizing $\Theta(p,q)$ is simply a minimization of a real-valued function of two variables, $p \in P$ and $q \in Q$. In many cases, the function $\Theta(p,q)$ is a measure of distance between p and q, such as $\Theta(p,q) = \|p-q\|_2^2$ for p and q in $\mathbb{R}^J$, or $\Theta(p,q) = KL(p,q)$, for non-negative vectors p and q. In the case of $\Theta(p,q) = \|p-q\|_2^2$, each step of the alternating minimization algorithm involves an orthogonal projection onto a closed and 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.

Let $f : \mathbb{R}^J \to \mathbb{R}$ be a Bregman function [21, 16, 22]; therefore f 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

$$D_f(x,z) := f(x) - f(z) - \langle \nabla f(z), x - z \rangle. \tag{14}$$

For example, the KL distance is a Bregman distance generated by the Bregman function

$$f(x) = \sum_{j=1}^{J} x_j \log x_j \ - x_j. \tag{15}$$

Suppose now that f is a Bregman function and P and Q are closed and convex subsets of the interior of the domain of f. Let p^{n+1} minimize $D_f(p,q^n)$ over all $p \in P$. It follows then that

$$\langle \nabla f(p^{n+1}) - \nabla f(q^n), p - p^{n+1} \rangle \geq 0, \tag{16}$$

for all $p \in P$. From the three-point identity of Chen and Teboulle [23] we have

$$D_f(p,q^n) - D_f(p^{n+1},q^n) = D_f(p,p^{n+1}) + \langle \nabla f(p^{n+1}) - \nabla f(q^n), p - p^{n+1} \rangle; \tag{17}$$

it follows that the three-point property holds, with

$$\Theta(p,q) = D_f(p,q), \tag{18}$$

and

$$\Delta(p,\hat{p}) = D_f(p,\tilde{p}). \tag{19}$$

To get the four-point property we need to restrict D_f somewhat; one such restriction is that $D_f(p,q)$ be jointly convex, that is, it be convex in the combined vector variable (p,q) [24]. The following lemma is due to Eggermont and LaRiccia [25].

Lemma 2.1 *Suppose that the Bregman distance $D_f(p,q)$ is jointly convex. Then it has the four-point property.*

The alternating minimization method works for any Bregman distance that is jointly convex. This includes the Euclidean and the KL distances.

3 Minimizing a Proximity Function

We present now an example of alternating Bregman distance minimization taken from [26]. The problem is the *convex feasibility problem* (CFP), to find a member of the intersection $C \subseteq \mathbb{R}^J$ of finitely many closed and convex sets C_i, $i = 1, ..., I$, or, failing that, to minimize the proximity function

$$F(x) = \sum_{i=1}^{I} D_i(\overleftarrow{P}_i x, x), \tag{20}$$

where f_i is a Bregman function 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 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

$$\Theta(p,q) = \sum_{i=1}^{I} D_i(c^i, x), \tag{21}$$

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

In [27], 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.

3.1 Right and Left Projections

Because Bregman distances D_f are not generally symmetric, we can speak of *right* and *left* Bregman projections onto a closed and 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 in (20) 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 and 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 [28] the authors note that the alternating minimization algorithm of [26] involves right and left Bregman projections, which suggests to them iterative methods involving a wider class of operators that they call "Bregman retractions".

4 The Bauschke-Combettes-Noll Problem

In [17] Bauschke, Combettes and Noll consider the following problem:

$$\text{minimize } \Theta(p, q) = \Lambda(p, q) := \phi(p) + \psi(q) + D_f(p, q), \tag{22}$$

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

$$b := \inf_{(p,q)} \Lambda(p, q) > -\infty, \tag{23}$$

and seek a sequence $\{(p^n, q^n)\}$ such that $\Lambda(p^n, q^n)$ converges to b. The sequence is obtained by the AM method, as in our previous discussion. They prove that, if the

Bregman distance is jointly convex, then $\{\Lambda(p^n, q^n)\} \downarrow b$. In this 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

$$\Lambda(p, q^{n-1}) - \Lambda(p^n, q^{n-1}) \geq \Lambda(p, q^n) - \Lambda(p, q). \tag{24}$$

A simple calculation shows that (24) is equivalent to

$$\Lambda(p, q) - \Lambda(p^n, q^n) \geq D(p, q^n) + D(p^n, q^{n-1}) - D(p, q^{n-1}) - D(p^n, q^n). \tag{25}$$

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

$$\Lambda(p, q) - \Lambda(p^n, q^n) \geq \langle \nabla_p \Lambda(p^n, q^n), p - p^n \rangle + \langle \nabla_q \Lambda(p^n, q^n), q - q^n \rangle, \tag{26}$$

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), and, similarly, $\nabla_q \Lambda(p_n, q_n)$.

Since q^n minimizes $\Lambda(p^n, q)$, it follows that

$$\nabla_q \Lambda(p^n, q^n) = 0, \tag{27}$$

for all q. Therefore,

$$\Lambda(p, q) - \Lambda(p^n, q^n) \geq \langle \nabla_p \Lambda(p^n, q^n), p - p^n \rangle . \tag{28}$$

We have

$$\langle \nabla_p \Lambda(p^n, q^n), p - p^n \rangle = \langle \nabla f(p^n) - \nabla f(q^n), p - p^n \rangle + \langle \nabla \phi(p^n), p - p^n \rangle. \tag{29}$$

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

$$\nabla_p \Lambda(p^n, q^{n-1}) = 0, \tag{30}$$

or

$$\nabla \phi(p^n) = \nabla f(q^{n-1}) - \nabla f(p^n), \tag{31}$$

so that

$$\langle \nabla_p \Lambda(p^n, q^n), p - p^n \rangle = \langle \nabla f(q^{n-1}) - \nabla f(q^n), p - p^n \rangle \tag{32}$$

$$= D(p, q^n) + D(p^n, q^{n-1}) - D(p, q^{n-1}) - D(p^n, q^n). \tag{33}$$

Using (28) we obtain (25). This shows that $\Lambda(p, q)$ has the five-point property whenever the Bregman distance $D = D_f$ is jointly convex.

From our previous discussion of AM, we conclude that the sequence $\{\Lambda(p^n, q^n)\}$ converges to b; this is Corollary 4.3 of [17].

In [29] it was shown that, in certain cases, the expectation maximization maximum likelihood (EM) method involves alternating minimization of a function of the form $\Lambda(p, q)$.

5 The SUMMA

We turn now to an apparently unrelated problem. Let S be an arbitrary set and $f : S \to (-\infty, \infty]$. The problem is to minimize f over a (not necessarily proper) subset C of S. At the nth step of a *sequential unconstrained minimization* algorithm, we obtain x^n by minimizing the function

$$G_n(x) = f(x) + g_n(x), \tag{34}$$

where the auxiliary function g_n is appropriately chosen [12]. If C is a proper subset of S we may force $g_n(x) = +\infty$ for x not in C, as in the barrier-function methods; then each x^n will lie in C.

The objective is to select the g_n so that the sequence $\{x^n\}$ converges to a solution of the problem, or failing that, at least to have the sequence $\{f(x^n)\}$ converging to the infimum of f over x in C.

In [13] we presented a particular class of sequential unconstrained minimization methods called SUMMA. As we showed in that paper, this class is broad enough to contain barrier-function methods, proximal minimization methods, the entropic proximal method of Teboulle [14], and the SMART. By reformulating the problem, penalty-function methods can also be shown to be members of the SUMMA class. When [13] was written, we were not able to include the EMML algorithm within the SUMMA class. As we shall see shortly, any AM problem with the five-point property can be reformulated as a SUMMA problem; therefore the EMML, which is such an AM algorithm, must also be a SUMMA algorithm.

For a method to be in the SUMMA class we require that $x^n \in C$, for each n, and that each auxiliary function g_n be finite for $x \in C$ and satisfy the inequalities

$$0 \leq g_{n+1}(x) \leq G_n(x) - G_n(x^n), \tag{35}$$

for all x. Note that it follows that $g_{n+1}(x^n) = 0$, for all n. We assume that $b := \inf_{x \in C} f(x) > -\infty$. The next two results are taken from [13].

Proposition 5.1 *The sequence* $\{f(x^n)\}$ *is non-increasing and the sequence* $\{g_n(x^n)\}$ *converges to zero.*

Theorem 5.1 *The sequence* $\{f(x^n)\}$ *converges to* b.

6 Examples of SUMMA

In this section we present several examples of SUMMA.

6.1 Barrier-Function Methods

Let $b : \mathbb{R}^J \to (-\infty, +\infty]$ be a continuous function, with effective domain

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

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

$$\text{minimize } f(x) + \frac{1}{n} b(x) \tag{36}$$

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

The iterative step of the barrier-function method can be formulated as follows:

$$\text{minimize } f(x) + [(n-1)f(x) + b(x)] \tag{37}$$

to get x^n. Since, for $n = 2, 3, ...$, the function $(n-1)f + b$ is minimized by x^{n-1}, the function

$$g_n(x) = (n-1)f(x) + b(x) - (n-1)f(x^{n-1}) - b(x^{n-1}) \tag{38}$$

is non-negative, and x^n minimizes the function $G_n = f + g_n$. From

$$G_n(x) = f(x) + (n-1)f(x) + b(x) - (n-1)f(x^{n-1}) - b(x^{n-1}), \tag{39}$$

it follows that

$$G_n(x) - G_n(x^n) = nf(x) + b(x) - nf(x^n) - b(x^n) = g_{n+1}(x), \tag{40}$$

so that g_{n+1} satisfies the condition in (35). This shows that the barrier-function method is a particular case of SUMMA.

6.2 Penalty-Function Methods

Once again, we want to minimize f over $x \in C$. In penalty-function methods the nth step is to minimize

$$f(x) + np(x), \tag{41}$$

where $p(x) > 0$ for x not in C and $p(x) = 0$ for $x \in C$. To show that penalty-function methods can be viewed as members of the SUMMA class, we reformulate these methods as barrier-function methods. In order to relate penalty-function methods to barrier-function methods, we note that minimizing $f + np$ is equivalent to minimizing $p + \frac{1}{n}f$. This is the form of the barrier-function iteration, with p now in the role previously played by f, and f now in the role previously played by b. We are not concerned here with the effective domain of f. See [13] for details.

6.3 Proximity-Function Minimization

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

$$D_h(x, z) = h(x) - h(z) - \langle \nabla h(z), x - z \rangle. \tag{42}$$

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

At the nth step of the *proximal minimization algorithm* (PMA) [30, 16], we minimize the function

$$G_n(x) = f(x) + D_h(x, x^{n-1}), \tag{43}$$

to get x^n. The function

$$g_n(x) = D_h(x, x^{n-1}) \tag{44}$$

is non-negative and $g_n(x^{n-1}) = 0$. We assume that each x^n lies in int D.

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

Lemma 6.1 *For each n we have*

$$G_n(x) = G_n(x^n) + D_f(x, x^n) + D_h(x, x^n). \tag{45}$$

Proof: Since x^n minimizes G_n within the set D, we have

$$0 = \nabla f(x^n) + \nabla h(x^n) - \nabla h(x^{n-1}). \tag{46}$$

Then

$$G_n(x) - G_n(x^n) = f(x) - f(x^n) + h(x) - h(x^n) - \langle \nabla h(x^{n-1}), x - x^n \rangle. \tag{47}$$

Now substitute, using (46) and the definition of Bregman distances. □

It follows from Lemma 6.1 that

$$G_n(x) - G_n(x^n) = g_{n+1}(x) + D_f(x, x^n). \tag{48}$$

7 AM as SUMMA

We show now that the SUMMA class of sequential unconstrained minimization algorithms includes all the AM methods for which the five-point property holds.

7.1 Reformulating AM as SUMMA

For each p in the set P, Let $f(p) = \Theta(p, q(p))$, where $q(p)$ is a member of Q for which $\Theta(p, q(p)) \leq \Theta(p, q)$, for all $q \in Q$.

At the nth step of AM we minimize

$$G_n(p) = \Theta(p, q^{n-1}) = \Theta(p, q(p)) + \Big(\Theta(p, q^{n-1}) - \Theta(p, q(p))\Big) \tag{49}$$

to get p^n. With

$$g_n(p) = \Big(\Theta(p, q^{n-1}) - \Theta(p, q(p))\Big) \geq 0, \tag{50}$$

we can write

$$G_n(p) = f(p) + g_n(p). \tag{51}$$

According to the five-point property, we have

$$G_n(p) - G_n(p^n) \geq \Theta(p, q^n) - \Theta(p, q(p)) = g_{n+1}(p). \tag{52}$$

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

We have seen that both the SMART and the EMML can be obtained as AM algorithms for which the five-point property holds. Consequently, both SMART and EMML are particular cases of SUMMA.

8 Conclusions

It was shown previously in [13] that the SUMMA class includes a wide variety of optimization algorithms, including the barrier-function methods, the proximal minimization algorithm of Censor and Zenios [15, 16], the entropic proximal method of Teboulle [14], and the simultaneous multiplicative algebraic reconstruction technique (SMART) [9, 10, 7, 8]. With some reformulation, it also contains the penalty-function methods. We have now shown that the alternating minimization methods of [1] are included in the SUMMA class whenever the five-point property holds. As a consequence, we learn that the EMML algorithm for Poisson mixtures [3, 4, 5, 6, 7, 8] is also a member of the SUMMA class.

References

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

2. Rockmore, A., Macovski, A.: A maximum likelihood approach to emission image reconstruction from projections. IEEE Transactions on Nuclear Science NS-23, 1428–1432 (1976)

3. Shepp, L., Vardi, Y.: Maximum likelihood reconstruction for emission tomography. IEEE Transactions on Medical Imaging MI-1, 113–122 (1982)

4. Lange, K., Carson, R.: EM reconstruction algorithms for emission and transmission tomography. Journal of Computer Assisted Tomography 8, 306–316 (1984)

5. Vardi, Y., Shepp, L.A., Kaufman, L.: A statistical model for positron emission tomography. Journal of the American Statistical Association 80, 8–20 (1985)

6. Lange, K., Bahn, M., Little, R.: A theoretical study of some maximum likelihood algorithms for emission and transmission tomography. IEEE Transactions on Medical Imaging MI-6(2), 106–114 (1987)

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

8. Byrne, C.: Erratum and addendum to 'Iterative image reconstruction algorithms based on cross-entropy minimization'. IEEE Transactions on Image Processing IP-4, 225–226 (1995)

9. Darroch, J., Ratcliff, D.: Generalized iterative scaling for log-linear models. Annals of Mathematical Statistics 43, 1470–1480 (1972)

10. Schmidlin, P.: Iterative separation of sections in tomographic scintigrams. Nuclear Medicine 9(1), 1–16 (1972)

11. Byrne, C.: Iterative reconstruction algorithms based on cross-entropy minimization. In: Levinson, S.E., Shepp, L. (eds.): Image Models (and their Speech Model Cousins), IMA Volumes in Mathematics and its Applications, vol. 80, pp. 1–11. Springer-Verlag, New York (1996)

12. Fiacco, A., McCormick, G.: Nonlinear Programming: Sequential Unconstrained Minimization Techniques. SIAM Classics in Mathematics, Philadelphia (1990)

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

14. Teboulle, M.: Entropic proximal mappings with applications to nonlinear programming. Mathematics of Operations Research 17(3), 670–690 (1992)

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

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

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

18. Kullback, S., Leibler, R.: On information and sufficiency. Annals of Mathematical Statistics 22, 79–86 (1951)

19. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B, 37, 1–38 (1977)

20. McLachlan, G.J., Krishnan, T.: The EM Algorithm and Extensions. John Wiley and Sons, Inc., New York (1997)

21. Bregman, L.M.: 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, 200–217 (1967)

22. Butnariu, D., Byrne, C., Censor, Y.: Redundant axioms in the definition of Bregman functions. Journal of Convex Analysis 10, 245–254 (2003)

23. Chen, G., Teboulle, M.: Convergence analysis of a proximal-like minimization algorithm using Bregman functions. SIAM Journal on Optimization 3, 538–543 (1993)

24. Bauschke, H., Borwein, J.: Joint and separate convexity of the Bregman distance. In: Butnariu, D., Censor, Y., Reich, S. (eds.): Inherently Parallel Algorithms in Feasibility and Optimization and their Applications, Studies in Computational Mathematics 8, pp. 23–36, Elsevier Publ., Amsterdam (2001)

25. Eggermont, P., LaRiccia, V.: Maximum Penalized Likelihood Estimation, Volume I: Density Estimation. Springer, New York (2001)

26. Byrne, C., Censor, Y.: Proximity function minimization using multiple Bregman projections, with applications to split feasibility and Kullback-Leibler distance minimization. Annals of Operations Research 105, 77–98 (2001)

27. Censor, Y., Elfving, T.: A multi-projection algorithm using Bregman projections in a product space. Numerical Algorithms 8, 221–239 (1994)

28. Bauschke, H., Combettes, P.: Iterating Bregman retractions. SIAM Journal on Optimization 13, 1159–1173 (2003)

29. Byrne, C., Eggermont, P.: EM Algorithms. in preparation (2012)

30. Byrne, C.: Bregman-Legendre multi-distance projection algorithms for convex feasibility and optimization. In: Butnariu, D., Censor, Y., Reich, S. (eds.): Inherently Parallel Algorithms in Feasibility and Optimization and their Applications, Studies in Computational Mathematics 8, pp. 87–100, Elsevier Publ., Amsterdam (2001)