AUXILIARY-FUNCTION MINIMIZATION ALGORITHMS

CHARLES L. BYRNE

Abstract. Let C be a nonempty subset of an arbitrary set X and $f : X \to \mathbb{R}$. The objective is to minimize $f(x)$ over $x \in C$. We get x^k, for $k = 1, 2, ...$, by minimizing $G_k(x) = f(x) + g_k(x)$ over all $x \in X$. We call this approach an *auxiliary-function* (AF) method if $g_k : X \to [0, +\infty]$, $g_k(x^{k-1}) = 0$, and $g_k(x) < +\infty$ if and only if $x \in C$. Then $\{f(x^k)\} \downarrow \beta^* \geq -\infty$. We consider conditions on the auxiliary functions g_k that guarantee that $\beta^* = \beta \doteq \inf_{x \in C} f(x)$.

An AF algorithm is said to be in the SUMMA class if the SUMMA Inequality, $G_k(x) - G_k(x^k) \geq g_{k+1}(x)$, for all $x \in X$, holds for all k, in which case it follows that $\beta^* = \beta$. We consider a variety of AF algorithms that either are in the SUMMA class or can be reformulated to be such. We also study some AF algorithms that are not in the SUMMA class, but for which $\beta^* = \beta$. This leads to a larger class, the SUMMA2 class of AF algorithms.

An AF algorithm is a proximal minimization algorithm (PMA) if $g_k(x) = d(x, x^{k-1})$, where $d : X \times X \to [0, +\infty]$ is a distance, so that $d(x, y) = 0$ if and only if $x = y$. Optimization transfer (OT) algorithms in statistics can be reformulated as PMA algorithms, as can the alternating-minimization (AM) algorithms of Csiszár and Tusnády. The "five-point property"(5PP) in AM, used by Csiszár and Tusnády to get $\beta^* = \beta$, is equivalent to the SUMMA Inequality, while the "weak"5PP (w5PP) implies membership in the SUMMA2 class.

1. Introduction

Let C be a nonempty subset of an arbitrary set X and $f : X \to \mathbb{R}$. We begin by considering the general problem of minimizing $f(x)$ over $x \in C$. Later in our discussion we shall let $C \subseteq X \doteq \mathbb{R}^N$ be a nonempty closed convex set. To enforce the restriction to the subset C we select *auxiliary functions* $g_k : X \to [0, +\infty]$ with $g_k(x)$ finite if and only if $x \in C$. We say that an iterative algorithm for minimizing $f(x)$ over $x \in C \subseteq X$ is an *auxiliary-function* (AF) method if $x^k \in C$ minimizes $G_k(x) = f(x) + g_k(x)$,

Date: October 27, 2017.

2010 *Mathematics Subject Classification.* Primary 65F10, 65K10; Secondary 90C26, 26B25.

Key words and phrases. Alternating minimization, optimization transfer, proximal minimization, Bregman distance, convex functions.

THIS PAPER IS DEDICATED TO MY FRIEND AND COLLEAGUE YAIR CENSOR ON THE OCCASION OF HIS 75TH BIRTHDAY.

where $g_k(x^{k-1}) = 0$. We can see easily that the sequence $\{f(x^k)\}$ is decreasing and converges to some $\beta^* \geq -\infty$. There need not be an $x \in C$ that minimizes $f(x)$; therefore we shall focus on conditions on the auxiliary functions $g_k(x)$ that guarantee that $\beta^* = \beta \doteq \inf_{x \in C} f(x)$. Auxiliary-function methods are similar to, but more general than, the sequential unconstrained minimization techniques treated in the classic text of Fiacco and McCormick [58]. In addition to enforcing restriction to a subset, as with barrier-function methods, auxiliary functions can be introduced to stabilize an ill-conditioned problem through regularization, to accelerate convergence to a solution, as sometimes happens with relaxation methods, or to simplify calculations.

1.1. **An Ill-conditioned Problem.** Let A be a real M by N matrix, with $M \geq N$ and A^TA invertible. If the ratio of the largest eigenvalue of A^TA to the smallest is much greater than one the problem of minimizing the function $f(x) = \frac{1}{2}\|Ax - b\|^2$ will be ill-conditioned. In that case, small changes in b can lead to large changes in the computed solution. Sometimes the norm of the computed solution will be unreasonably large. This happens in band-limited extrapolation [36], but, somewhat surprisingly, the instability can be helpful in solving the optical phase retrieval problem [15]. To regularize the problem and control the growth of the norm we can minimize $f(x) + \frac{1}{k}\|x\|^2$ to obtain the approximate solution x^k. As $k \to +\infty$ the sequence $\{f(x^k)\} \downarrow \inf f(x)$. With additional conditions we can have convergence of $\{x^k\}$ to a minimizer of $f(x)$. However, as pointed out in [46], as $k \to +\infty$ the constrained problem becomes increasingly as ill-conditioned as the unconstrained problem.

1.2. **Barrier-function Methods.** The barrier-function approach is a good illustration of the use of AF algorithms for constrained minimization. Suppose that X is an arbitrary set, $C \subseteq X$ a nonempty subset, $f : X \to \mathbb{R}$, and we want to minimize $f(x)$ over $x \in C$. Using the barrier-function approach, we select a function $b : X \to (0, +\infty]$, with $b(x) < +\infty$ if and only if $x \in C$, and minimize $B_k(x) = f(x) + \frac{1}{k}b(x)$ to get $x^k \in C$. If $C = X$, then this is regularization. We have the following theorem.

Theorem 1.1. *The sequence $\{f(x^k)\}$ is decreasing to a limit $\beta^* \geq \beta \doteq \inf_{x \in C} f(x)$, the sequence $\{b(x^k)\}$ is increasing, and $\beta^* = \beta$.*

Proof. Since $B_k(x^k) \leq B_k(x^{k-1})$ and $B_{k-1}(x^{k-1}) \leq B_{k-1}(x^k)$, we have

$$\frac{1}{k-1}[b(x^{k-1}) - b(x^k)] \geq f(x^{k-1}) - f(x^k) \geq \frac{1}{k}[b(x^{k-1}) - b(x^k)],$$

establishing the first two claims in the theorem. Now suppose that $\beta^* > \beta$. Then there must be $z \in C$ with $f(x^k) \geq \beta^* > f(z) \geq \beta$, for all k. From $B_k(z) \geq B_k(x^k)$ we get

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

But $\frac{1}{k}\left(b(z) - b(x^k)\right) \to 0$, since $\frac{1}{k}b(z) \downarrow 0$. □

It is helpful to note that minimizing $B_k(x)$ is equivalent to minimizing

$$kf(x)+b(x)=f(x)+(k-1)f(x)+b(x)=f(x)+(k-1)B_{k-1}(x)$$

and therefore x^k minimizes

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

with $d_k(x,x^{k-1})\geq 0$ and $d_k(x^{k-1},x^{k-1})=0$. This is something like a proximal minimization algorithm to be discussed shortly, except that the distances d_k vary with k and depend on the function $f(x)$. With $g_k(x)=d_k(x,x^{k-1})$ above we have an AF algorithm and

$$G_k(x)-G_k(x^k)=g_{k+1}(x), \tag{1.1}$$

which will serve to motivate our definition of the SUMMA class of AF algorithms. A penalty-function method for minimizing $f(x)$ over $x\in C$ is to minimize $f(x)+kp(x)$ to get x^k, where $p:X\to[0,+\infty)$ and $p(x)=0$ if and only if $x\in C$. Writing $f(x)+kp(x)$ as $p(x)+\frac{1}{k}f(x)$, we find that penalty-function methods can be analyzed using the barrier-function approach [35]. For further discussion of barrier-function and penalty-function methods and related ideas see [58].

1.3. **Proximal Minimization Algorithms.** Once again, however, the constrained problem of minimizing $B_k(x)$ may grow increasingly ill-conditioned as k increases. Censor and Zenios [46] suggest that we consider relaxation methods to minimize $f(x)$ over $x\in C$. General proximal minimization algorithms (PMA) are a type of relaxation algorithms in which we minimize $f(x)+d(x,x^{k-1})$ to get x^k, where $f:X\to\mathbb{R}$, $d(x,y)\geq 0$ and $d(x,y)=0$ if and only if $x=y$. Clearly the sequence $\{f(x^k)\}$ is decreasing to some $\beta^*\geq\beta\doteq\inf_{x\in C}f(x)$. Again, we want $\beta^*=\beta$. With additional conditions placed on X, f and d we can say more, as we shall see.

1.4. **Using Bregman Distances.** The methods called proximal minimization using D-functions (PMD) in [46], and called in this paper PMAB methods, involve minimizing $G_k(x)\doteq f(x)+D_h(x,x^{k-1})$ to get x^k, where $X=\mathbb{R}^N$, $D_h(x,y)$ is a Bregman distance [10, 6, 46, 24, 13], with

$$D_h(x,y)\doteq h(x)-h(y)-\langle\nabla h(y),x-y\rangle \tag{1.2}$$

and $f:X\to\mathbb{R}$ is convex. The reader should note that we use the term *Bregman distance* in a somewhat looser sense than in [46, 13] and elsewhere. For us a generalized distance $D_h(x,z)$ will be called a Bregman distance if it has the form given in Equation (1.2), where $h:C\subseteq\mathbb{R}^N\to\mathbb{R}$ is convex on the closed convex set C and differentiable in the nonempty interior of C. We will assume also that, for each k, x^k is the unique minimizer of $f(x)+D_h(x,x^{k-1})$ and lies in the interior of C.

To prove convergence of the PMAB algorithm we will need additional assumptions. As we shall see shortly, all PMAB algorithms are in the SUMMA class, since

$$G_k(x) - G_k(x^k) \geq D_h(x, x^k) \tag{1.3}$$

for all x. Therefore, the sequence $\{f(x^k)\} \downarrow \beta = \inf_{x \in C} f(x)$. From the inequality in (1.3) we have

$$D_h(x, x^{k-1}) - D_h(x, x^k) \geq f(x^k) - f(x), \tag{1.4}$$

for all x. If there is $\hat{x} \in C$ such that $f(x) \geq f(\hat{x})$, for all $x \in C$, then

$$D_h(\hat{x}, x^{k-1}) - D_h(\hat{x}, x^k) \geq f(x^k) - f(\hat{x}) \geq 0, \tag{1.5}$$

for all k. Therefore, the sequence $\{D_h(\hat{x}, x^k)\}$ is decreasing. If the Bregman distance $D_h(z, \cdot)$ has bounded level sets, then the sequence $\{x^k\}$ is bounded, there is a cluster point of the sequence, call it x^*, and $f(x^*) = f(\hat{x})$. Replacing $\hat{x}$ with x^*, we find that the sequence $\{D_h(x^*, x^k)\}$ is decreasing. Under reasonable assumptions on D_h [46, 35] it will follow that a subsequence converges to zero, the entire sequence converges to zero, and the sequence $\{x^k\}$ converges to x^*.

2. The SUMMA Class

An AF algorithm is said to belong to the SUMMA class if the following SUMMA Inequality holds for all k and $x \in X$:

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

We already know that $\{f(x^k)\} \downarrow \beta^* \geq \beta \doteq \inf_{x \in C} f(x)$. We have the following theorem.

Theorem 2.1. *If an AF algorithm is in the SUMMA class then $\beta^* = \beta$.*

Proof. From the inequality in (2.1) we have

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

If $\beta^* > \beta$ then there is $z \in C$ with

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

so that

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

But the decreasing sequence $\{g_k(z)\}$ cannot have successive increments bounded away from zero. □

From Equation (1.1) we see that the barrier-function algorithms are in the SUMMA class. Proximal minimization algorithms using Bregman distances (PMAB) are also in the SUMMA class.

Theorem 2.2. *Let $f : \mathbb{R}^N \to (-\infty, +\infty]$ be convex and, for each k, $g_k(x) = D_h(x, x^{k-1})$. Then the AF algorithm is in the SUMMA class.*

Proof. Since x^k minimizes $f(x) + h(x) - h(x^{k-1}) - \langle \nabla h(x^{k-1}), x - x^{k-1}\rangle$ it follows that

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

where $\partial f(x^k)$ denotes the subdifferential of f at x^k. Therefore, there is $u^k \in \partial f(x^k)$ with

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

From

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

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

it follows that

$$G_k(x) - G_k(x^k) = f(x) - f(x^k) - \langle u^k, x - x^k\rangle + D_h(x, x^k)$$

and so

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

□

Corollary 2.3. *If $g_k(x) = \frac{1}{2}\|x - x^{k-1}\|^2$ then the AF algorithm to minimize the convex function $f(x)$ over all $x \in \mathbb{R}^N$ is PMAB and therefore is in the SUMMA class.*

For $a > 0$ and $b > 0$, $KL(a, b) = a \log \frac{a}{b} + b - a$ is the Kullback-Leibler distance [64], which is positive, unless $a = b$. Using limits, we define $KL(a, 0) = +\infty$ and $KL(0, b) = b$. We then extend coordinate-wise to get

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

for nonnegative vectors x and z. With $x_+ \doteq \sum_{j=1}^{J} x_j$ we have the identity

$$KL(x, z) = KL(x_+, z_+) + KL(x, \frac{x_+}{z_+} z) \qquad (2.2)$$

from which we get the useful inequality

$$KL(x_+, z_+) \leq KL(x, z). \qquad (2.3)$$

The KL distance is a Bregman distance D_h for $h(x) = \sum_{j=1}^{J} x_j \log x_j - x_j$.

Corollary 2.4. *If $g_k(x) = KL(x, x^{k-1})$, for nonnegative vectors x and x^{k-1}, then the AF algorithm to minimize the convex function $f(x)$ over all $x \geq 0$ is PMAB and therefore is in the SUMMA class.*

3. The EM Algorithm and PMA

Maximizing the likelihood function is a well known tool in statistical parameter estimation. Assume that Y is a random vector governed by a probability density function or probability function $f_Y(y|\theta^*)$, for some parameter vector $\theta^* \in \Theta$. We have one realization y of Y, from which we want to estimate θ^*. Our maximum-likelihood estimate is the θ for which the likelihood function $L(\theta) \doteq f_Y(y|\theta)$ is maximized over $\theta \in \Theta$. The *expectation maximization* (EM) algorithm [52, 67] is not one algorithm, but a template or recipe for the design of iterative methods for maximizing likelihood in statistics. As discussed in [35], the usual presentation of the EM algorithm, as found in [67] and elsewhere, is flawed. The STEM approach discussed here can be viewed as an improvement upon the usual EM method, although the two are the same in most cases.

In this section we present our nonstochastic EM for optimization and define our STEM template in terms of NSEM. It will follow from results concerning NSEM that likelihood is always increasing for STEM algorithms.

3.1. **NSEM.**

We assume that there is a function $b : \Theta \times \Omega \to \mathbb{R}_+$, where (Ω, μ) is a measure space and

$$a(\theta) \doteq \int_\Omega b(\theta, \omega) d\mu(\omega). \tag{3.1}$$

Let $f(\theta) = -a(\theta)$ and θ^0 be arbitrary. For $k = 1, 2, ...$, we maximize

$$\int_\Omega b(\theta^{k-1}, \omega) \log b(\theta, \omega) d\mu(\omega) \tag{3.2}$$

to get θ^k. Note that the integration may be replaced by summation, as needed. Using the Kullback–Leibler distance, we can reformulate the NSEM.

With the shorthand notation $b(\theta) = b(\theta, \omega)$ we define

$$KL\left(b(\theta), b(\gamma)\right) = \int_\Omega KL\left(b(\theta, \omega), b(\gamma, \omega)\right) d\mu(\omega).$$

Proposition 3.1. *The sequence $\{a(\theta^k)\}$ is increasing.*

Proof. We have

$$a(\theta^{k-1}) = a(\theta^{k-1}) - KL\left(b(\theta^{k-1}), b(\theta^{k-1})\right) \leq a(\theta^k) - KL\left(b(\theta^{k-1}), b(\theta^k)\right).$$

Therefore,

$$a(\theta^k) - a(\theta^{k-1}) \geq KL\left(b(\theta^{k-1}), b(\theta^k)\right).$$

□

We see easily that θ^k minimizes

$$G_k(\theta) = KL\left(b(\theta^{k-1}), b(\theta)\right) - a(\theta) = f(\theta) + d(\theta, \theta^{k-1}), \tag{3.3}$$

for

$$d(\theta, \gamma) = KL\left(b(\gamma), b(\theta)\right).$$

Consequently, the NSEM is a PMA.

3.2. **STEM.** Now we define the STEM class of iterative algorithms as a subclass of the NSEM. For any random vectors X and Y governed by the joint probability density function or joint probability function $f_{X,Y}(x,y|\theta)$ we have

$$f_Y(y|\theta) = \int f_{X,Y}(x,y|\theta)dx. \tag{3.4}$$

With $a(\theta) = f_Y(y|\theta)$ and $b(\theta,\omega) = f_{X,Y}(x,y|\theta)$ we see that Equation (3.4) becomes Equation (3.1). For the case of probability functions, the integration is replaced by summation. So our STEM template fits into that of the NSEM. The iterative step is then to find θ^k by maximizing the function

$$\int f_{X,Y}(x,y|\theta^{k-1}) \log f_{X,Y}(x,y|\theta)dx.$$

It follows from our discussion of the NSEM that the sequence $\{f_Y(y|\theta^k)\}$ is increasing. Of course, additional restrictions are needed to prove that the sequence $\{\theta^k\}$ converges to a maximizer of the likelihood function $L(\theta) = f_Y(y|\theta)$.

4. Concerning Computation

We haven't said anything yet about the difficulties involved in computing the x^k in AF algorithms. Minimizing $f(x) + \frac{1}{2}\|x - x^{k-1}\|^2$ leads to

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

so we do not have a closed-form expression for x^k. Similarly, minimizing $f(x) + KL(x, x^{k-1})$ over $x \geq 0$ leads to

$$\log x_j^k = \log x_j^{k-1} - \nabla f(x^k)_j.$$

Once again, we have no closed-form expression for x^k. How can we remedy this situation?

4.1. **A Remedy.** Suppose that we select $g(x)$ convex and differentiable, with $h(x) \doteq g(x) - f(x)$ also convex. Then minimizing $f(x) + D_h(x, x^{k-1})$ is equivalent to minimizing $f(x) + D_g(x, x^{k-1}) - D_f(x, x^{k-1})$. Therefore, we have

$$\nabla g(x^k) = \nabla g(x^{k-1}) - \nabla f(x^{k-1}).$$

Now suppose we have selected $g(x)$ so that this equation is easily solved. Then we would have a closed-form expression for the iterate x^k.

For example, suppose that $\nabla f(x)$ is L-Lipschitz continuous and $0 < \gamma < \frac{1}{L}$. Then $h(x) \doteq \frac{1}{2\gamma}\|x\|^2 - f(x)$ is convex. Minimizing $f(x) + D_h(x, x^{k-1})$ now leads to

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

which is a gradient-descent algorithm. If $f(x) = \frac{1}{2}\|Ax - b\|^2$ we get the Landweber algorithm [65].

The *simultaneous multiplicative algebraic reconstruction technique*, the SMART [51, 74, 44, 17, 19], is an iterative algorithm that minimizes $f(x) = KL(Px, y)$ over $x \geq 0$, where P is an I by J matrix with nonnegative entries and y is a positive vector. If we enforce the nonnegativity constraint by minimizing $KL(Px, y) + KL(x, x^{k-1})$ to get x^k we do not obtain x^k in closed form. However, if the column sums of the matrix P are all equal to one, then $KL(Px, Px^{k-1}) = D_f(x, x^{k-1})$ and $D_h(x, x^{k-1}) \doteq KL(x, x^{k-1}) - KL(Px, Px^{k-1})$ is a Bregman distance. The SMART iterative step, obtained by minimizing $KL(Px, y) + D_h(x, x^{k-1})$, is

$$x_j^k = x_j^{k-1} \exp\left(\sum_{i=1}^{I} P_{i,j} \log \frac{y_i}{Px_i^{k-1}}\right). \tag{4.1}$$

We shall discuss the SMART in more detail later in this paper.

4.2. **Forward-Backward Splitting.** The *forward-backward splitting* (FBS) algorithm [49] is used to minimize $f(x) = f_1(x) + f_2(x)$, where both f_1 and f_2 are convex, but f_1 need not be differentiable. When $\nabla f_2(x)$ is L-Lipschitz continuous and $0 < \gamma < \frac{1}{L}$ we have a Bregman distance

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

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

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

or

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

It follows from a characterization of Moreau's proximity operator prox [68, 69, 70, 49] that

$$x^k = \mathrm{prox}_{\gamma f_1}\left(x^{k-1} - \gamma \nabla f_2(x^{k-1})\right). \tag{4.2}$$

This is the FBS iterative step. The FBS is a PMAB algorithm, so we know that the sequence $\{f(x^k)\}$ is decreasing to $\beta \doteq \inf_x f(x)$. If $f_2(x) = \frac{1}{2}\|Ax - b\|^2$ and $f_1(x) = \iota_C(x)$, the function equal to zero for $x \in C$ and equal to $+\infty$ otherwise, we get the projected Landweber algorithm. We have the following convergence theorem for the FBS algorithm.

Theorem 4.1. *The sequence $\{x^k\}$ given by Equation (4.2) converges to a minimizer of the function $f(x) = f_1(x) + f_2(x)$, whenever minimizers exist.*

Proof. A relatively simple calculation shows that

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

$$\left(f_1(x) - f_1(x^k) - \frac{1}{\gamma}\langle (x^{k-1} - \gamma \nabla f_2(x^{k-1})) - x^k, x - x^k\rangle\right). \tag{4.3}$$

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,

$$G_k(x) - G_k(x^k) \geq \frac{1}{2\gamma}\|x - x^k\|_2^2 \geq g_{k+1}(x); \tag{4.4}$$

the SUMMA Inequality holds and the FBS algorithm is in 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 (4.4), 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. □

As shown in [49], using the Baillon–Haddad Theorem [3, 9, 37] and the theory of firmly non-expansive operators we can allow $0 < \gamma < \frac{2}{\rho(A^TA)}$.

4.3. The Split Feasibility Problem.

We apply the FBS algorithm to solve the *split feasibility problem* (SFP) [40, 24]: given a real M by N matrix A, a closed convex set $C \subseteq \mathbb{R}^N$ and a closed convex set $Q \subseteq \mathbb{R}^M$, find $x \in C$ with $Ax \in Q$. We consider the more general problem of minimizing the convex differentiable function $f_2(x) = \frac{1}{2}\|P_QAx - Ax\|^2$ over $x \in C$, where P_Q denotes the orthogonal projection onto the set Q. With $f_1(x) = \iota_C(x)$ we know $\text{prox}_{\gamma f_1}(x) = P_C(x)$, and the gradient of $f_2(x)$ is

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

For $0 < \gamma < \frac{1}{\rho(A^TA)}$ the iterative sequence

$$x^k = P_C\left(x^{k-1} - \gamma A^T(I - P_Q)Ax^{k-1}\right) \tag{4.5}$$

converges to a minimizer of $f(x)$ over $x \in C$, whenever such minimizers exist [27]. In recent work Yair Censor and his colleagues have generalized the CQ algorithm and applied it to problems in proton-beam and x-ray radiation therapy [42, 43, 72]. If $A = I$, the identity matrix, and $\gamma = 1$, then the iteration in Equation (4.5) becomes $x^k = P_CP_Qx^{k-1}$, which is the alternating orthogonal projection algorithm investigated in [47]. It is also an example of an alternating minimization algorithm, which we discuss later in this paper.

5. The SUMMA2 Class

We turn now to several AF algorithms that are not in the SUMMA class, but for which $\{f(x^k)\} \downarrow \beta^* = \beta \doteq \inf_{x \in C} f(x)$.

5.1. **Defining the SUMMA2 Class.**

We say that an AF method for minimizing $f(x)$ over $x \in C$ is in the SUMMA2 class if, for each sequence generated by the algorithm, there are functions $h_k : C \to \mathbb{R}_+$ such that

$$h_k(x) + f(x) \geq h_{k+1}(x) + f(x^k), \tag{5.1}$$

for all $x \in C$. We have the following theorem.

Theorem 5.1. *If an AF algorithm is in the SUMMA2 class, then $\beta^* = \beta$.*

Proof. If $\{f(x^k)\} \downarrow \beta^* > \beta \doteq \inf_{x \in C} f(x)$ then there is $z \in C$ with $\beta^* > f(z) \geq \beta$. Consequently, we have

$$h_k(z) - h_{k+1}(z) \geq f(x^k) - f(z) \geq \beta^* - f(z) > 0,$$

for all k, which cannot happen. □

5.2. **The Approach of Auslender and Teboulle.**

The method of Auslender and Teboulle [2] is a particular example of an AF algorithm not in the SUMMA class, but for which $\beta^* = \beta$. We take C to be a closed, convex subset of $\mathbb{R}^N$, with nonempty interior U. At the kth step of their method one minimizes a function

$$G_k(x) = f(x) + d(x, x^{k-1}) \tag{5.2}$$

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* $p(a, b) \geq 0$, finite for a and b in U, with $p(a, a) = 0$ and

$$\langle \nabla_1 d(b, a), c - b\rangle \leq p(c, a) - p(c, b), \tag{5.3}$$

for all c in U. They show that, if the distance d has associated with it an induced proximal distance, then $\beta^* = \beta$.

They consider two types of distances d for which there are induced proximal distances p: the first type are the Bregman distances $d = D_h$, which are self-proximal in the sense that $d = p$; the second type are those having the form

$$d(x,z) = d_\phi(x,z) \doteq \sum_{n=1}^{N} z_n \phi(\frac{x_n}{z_n}),$$

for functions ϕ having certain properties to be discussed below. In such cases the induced proximal distance is $p(x,z) = \phi^{''}(1)KL(x,z)$, where $KL(x,z)$ is the Kullback–Leibler distance. Then for all $x \geq 0$ we have

$$\phi^{''}(1)\left(KL(x,x^k) - KL(x,x^{k+1})\right) \geq f(x^k) - f(\hat{x}). \tag{5.4}$$

The Hellinger distance,

$$H(x,z) = \sum_{n=1}^{N} \left(\sqrt{x_n} - \sqrt{z_n}\right)^2,$$

fits into this framework.

The required conditions on the function $\phi(t)$ are as follows: $\phi : \mathbb{R} \to (-\infty, +\infty]$ is lower semi-continuous, proper and convex, with dom $\phi \subseteq \mathbb{R}_+$, and dom $\partial\phi = \mathbb{R}_{++}$. In addition, the function ϕ is C^2, strictly convex, and nonnegative on $\mathbb{R}_{++}$, with $\phi(1) = \phi'(1) = 0$, and

$$\phi''(1)\left(1 - \frac{1}{t}\right) \leq \phi'(t) \leq \phi''(1)\log(t). \tag{5.5}$$

For the Hellinger case we have $\phi(t) = (\sqrt{t} - 1)^2$, so that these conditions are satisfied and for all $x \geq 0$ we have

$$KL(x,x^k) - KL(x,x^{k+1}) \geq 2\left(f(x^k) - f(x)\right). \tag{5.6}$$

It can be shown that, whenever there is an induced proximal distance, then, for any x, we have

$$p(x,x^k) - p(x,x^{k+1}) \geq f(x^k) - f(x). \tag{5.7}$$

With $h_k(x) \doteq p(x,x^k)$, the algorithm falls into the SUMMA2 class, and so $\beta^* = \beta$.

5.3. **The EMML Algorithm.** The *expectation maximization maximum likelihood* (EMML) algorithm [77, 19] is an iterative algorithm that minimizes $f(x) = KL(y, Px)$ over $x \geq 0$. The iterative step of the EMML, similar to that in Equation (4.1), is

$$x_j^k = x_j^{k-1}\left(\sum_{i=1}^{I} P_{i,j}\left(\frac{y_i}{Px_i^{k-1}}\right)\right). \tag{5.8}$$

The EMML algorithm is not PMAB, and not in the SUMMA class, although it does minimize $f(x)$ over $x \geq 0$. As we shall see, the reason that $\beta^* = \beta$

here is that the EMML algorithm is in the broader SUMMA2 class of AF methods.

6. Alternating Minimization

In this section we review the basics of alternating minimization [50], and then show that AM and PMA are equivalent. Alternating minimization plays an important role in the application of the EM algorithm [52] to medical image reconstruction [75, 77, 19].

Proximal minimization algorithms (PMA), alternating minimization methods (AM), and optimization transfer (OT) are three well studied areas involving iterative minimization algorithms. Optimization transfer, also called *surrogate-function methods* or *majorization minimization*, commonly used in statistics [1, 66, 48] (see also [53]), uses $g(x|z) \geq f(x) = g(x|x)$ and x^k is obtained by minimizing $g(x|x^{k-1})$. With $d(x,z) \doteq g(x|z) - f(x)$, it is clear that the OT iteration is equivalent to minimizing $f(x) + d(x, x^{k-1})$, which shows that OT methods are equivalent to PMA. Alternating minimization methods are also equivalent to PMA, although showing this takes a bit more work [33].

6.1. **AM Algorithms are PMA.**

Let $\Phi : P \times Q \to (-\infty, +\infty]$, where P and Q are arbitrary nonempty sets. In the AM approach we minimize $\Phi(p, q^{k-1})$ over $p \in P$ to get p^k and then minimize $\Phi(p^k, q)$ over $q \in Q$ to get q^k. It follows immediately that the sequence $\{\Phi(p^k, q^k)\}$ is decreasing. We want

$$\{\Phi(p^k, q^k)\} \downarrow \beta \doteq \inf\{\Phi(p,q) | p \in P, q \in Q\}. \tag{6.1}$$

For each p select $q(p)$ for which $\Phi(p, q(p)) \leq \Phi(p, q)$ for all $q \in Q$. Then define $f(p) \doteq \Phi(p, q(p))$. Since $q^{k-1} = q(p^{k-1})$, we have

$$\Phi(p, q^{k-1}) = \Phi(p, q(p^{k-1})).$$

Minimizing $\Phi(p, q^{k-1})$ to get p^k is equivalent to minimizing

$$G_k(p) = \Phi(p, q(p)) + \Phi(p, q(p^{k-1})) - \Phi(p, q(p)) = f(p) + g_k(p), \tag{6.2}$$

where

$$g_k(p) = \Phi(p, q(p^{k-1})) - \Phi(p, q(p)).$$

Clearly, $g_k(p) \geq 0$ and $g_k(p^{k-1}) = 0$. Therefore, every AM algorithm is also an AF algorithm.

We define a "distance" $d(p, p')$ on the set $P \times P$ by

$$d(p, p') \doteq \Phi(p, q(p')) - \Phi(p, q(p)). \tag{6.3}$$

Then we see that p^k is obtained by minimizing $f(p) + d(p, p^{k-1})$. Consequently, every AM algorithm is PMA. The converse is obvious: minimizing $\Phi(x, x^{k-1}) = f(x) + d(x, x^{k-1})$ with respect to x gives $x = x^k$ and minimizing $\Phi(x^k, x) = f(x^k) + d(x^k, x)$ gives $x = x^k$ again. We can formulate an AM algorithm as OT by choosing $\Phi(p, q(p'))$ to play the role of $g(x|z)$ in OT.

6.2. **The Five-Point Property.** In [50] Csiszár and Tusnády show that, if the function Φ possesses what they call the *five-point property* (5PP),

$$\Phi(p,q)+\Phi(p,q^{k-1})\geq\Phi(p,q^k)+\Phi(p^k,q^{k-1}), \tag{6.4}$$

for all p, q, and k, then (6.1) 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, when the AM method is reformulated as above, as an AF method to minimize a function of the single variable p, the five-point property for AM is precisely the SUMMA Inequality.

It is often the case that AM methods are described using the *three-* and *four-point properties* (3PP and 4PP). The 3PP is

$$\Phi(p,q^{k-1})-\Phi(p^k,q^{k-1})\geq\Delta(p,p^k)\geq 0, \tag{6.5}$$

where $\Delta : P\times P\rightarrow\mathbb{R}_+$ and $\Delta(p,p)=0$, for all $p\in P$. The 4PP is the following:

$$\Delta(p,p^k)\geq\Phi(p,q^k)-\Phi(p,q), \tag{6.6}$$

for all p, q, and k. Clearly, the 3PP and 4PP together imply the 5PP.

When the 3PP and 4PP hold we have

$$\Delta(p,p')\geq d(p,p')=\Phi(p,q(p'))-\Phi(p,q(p)).$$

If we redefine Δ by $\Delta(p,p')\doteq d(p,p')$, then the 4PP is automatically true and the 3PP becomes equivalent to the 5PP. The 3PP is now

$$\Phi(p,q^{k-1})-\Phi(p^k,q^{k-1})\geq d(p,p^k). \tag{6.7}$$

The weak 3PP (w3PP), defined by

$$\Phi(p,q^{k-1})-\Phi(p^k,q^k)\geq d(p,p^k), \tag{6.8}$$

implies that the algorithm is in the SUMMA2 class, so is sufficient to guarantee that $\beta^*=\beta$.

It is shown in [56] that a Bregman distance that is jointly convex enjoys the 5PP with respect to closed, convex sets P and Q. Therefore $\Phi(p,q)=\frac{1}{2}\|p-q\|^2$ and $\Phi(p,q)=KL(p,q)$ both have the 5PP for appropriately defined P and Q. Related work is found in [8].

7. The SMART and the EMML Algorithms

In this section we present the tandem development of the SMART and the EMML algorithms, as originally published in [19]. We assume that y is a positive vector in $\mathbb{R}^I$, P an I by J matrix with nonnegative entries $P_{i,j}$, $s_j=\sum_{i=1}^I P_{i,j}>0$, and we want to find a nonnegative solution or approximate solution x for the linear system of equations $y=Px$. The EMML algorithm will minimize $KL(y,Px)$, while the SMART will minimize $KL(Px,y)$, over $x\geq 0$. For notational simplicity we shall assume that the system has been normalized so that $s_j=1$ for each j.

7.1. **The SMART.** The SMART algorithm [51, 74, 44, 17, 19] minimizes the function $f(x) = KL(Px, y)$, over nonnegative vectors x. Having found the vector x^{k-1}, the next vector in the SMART sequence is x^k, with entries given by

$$x_j^k = x_j^{k-1} \exp\left(\sum_{i=1}^{I} P_{ij} \log\left(\frac{y_i}{(Px^{k-1})_i}\right)\right). \tag{7.1}$$

The iterative step of the SMART can be described as $x^k = Sx^{k-1}$, where S is the operator defined by

$$(Sx)_j = x_j \exp\left(\sum_{i=1}^{I} P_{ij} \log\left(\frac{y_i}{(Px)_i}\right)\right). \tag{7.2}$$

In our proof of convergence of the SMART we will show that any cluster point x^* of the SMART sequence $\{x^k\}$ is a fixed point of the operator S. To avoid pathological cases in which $Px_i^* = 0$ for some index i, we can assume, at the outset, that all the entries of P are positive. This is wise, in any case, since the model of $y = Px$ is unlikely to be exactly accurate in applications.

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

$$x_j^k = x_j^{k-1} \left(\sum_{i=1}^{I} P_{ij} \left(\frac{y_i}{(Px^{k-1})_i}\right)\right). \tag{7.3}$$

The iterative step of the EMML algorithm can be described as $x^k = Mx^{k-1}$, where M is the operator defined by

$$(Mx)_j = x_j \left(\sum_{i=1}^{I} P_{ij} \left(\frac{y_i}{(Px)_i}\right)\right). \tag{7.4}$$

As we shall see, the EMML algorithm forces the sequence $\{KL(y, Px^k)\}$ to be decreasing. It follows that $(Px^*)_i > 0$, for any cluster point x^* and for all i.

7.3. **The SMART as AM.** In [17] the SMART was derived using the following alternating minimization (AM) approach. Let $\mathbb{X}$ be the set of all nonnegative x for which Px has only positive entries; all positive x are in $\mathbb{X}$.

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

$$r(x)_{ij} = x_j P_{ij} \left(\frac{y_i}{(Px)_i}\right), \tag{7.5}$$

and

$$q(x)_{ij} = x_j P_{ij}. \tag{7.6}$$

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

$$G_k(x) = KL(q(x), r(x^{k-1})) = \sum_{i=1}^{I}\sum_{j=1}^{J} KL(q(x)_{ij}, r(x^{k-1})_{ij}) \tag{7.7}$$

over $x \geq 0$. Note that $f(x) = KL(Px, y) = KL(q(x), r(x))$. We have the following helpful *Pythagorean identities*:

$$KL(q(x), r(z)) = KL(q(x), r(x)) + KL(x, z) - KL(Px, Pz); \tag{7.8}$$

and

$$KL(q(x), r(z)) = KL(q(Sz), r(z)) + KL(x, Sz). \tag{7.9}$$

Note that it follows from Equation (2.3) that $KL(x, z) - KL(Px, Pz) \geq 0$.

From the Pythagorean identities we find that x^k is obtained by minimizing

$$G_k(x) = KL\left(q(x), r(x^{k-1})\right) =$$

$$KL(Px, y) + KL(x, x^{k-1}) - KL(Px, Px^{k-1}), \tag{7.10}$$

so that

$$g_k(x) = KL(x, x^{k-1}) - KL(Px, Px^{k-1}). \tag{7.11}$$

Then

$$G_k(x) - G_k(x^k) = KL(x, x^k) \geq KL(x, x^k) - KL(Px, Px^k) = g_{k+1}(x).$$

Therefore, the SMART is in the SUMMA class. It follows from the SUMMA Inequality that, for all $x \geq 0$,

$$g_k(x) + f(x) \geq g_{k+1}(x) + f(x^k). \tag{7.12}$$

Since

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

the sequence $\{x^k\}$ is bounded and has a cluster point, x^*, with $f(x^k) \geq f(x^*)$ for all k. With $x = x^*$ in (7.12), we obtain

$$D_h(x^*, x^{k-1}) - D_h(x^*, x^k) \geq f(x^k) - f(x^*) \geq 0.$$

Therefore, the sequence $\{f(x^k)\}$ converges to $f(x^*)$. Since the SMART is in SUMMA, we know that $f(x^*)$ must be the minimum of $f(x)$. Since a subsequence of $\{D_h(x^*, x^k)\}$ converges to zero, it follows that $\{x^k\}$ converges to x^*.

Let $\hat{x}$ be any minimizer of $KL(Px, y)$. Using the Pythagorean identites we find that

$$\begin{aligned} KL(\hat{x}, x^k) - KL(\hat{x}, x^{k+1}) = KL(Px^{k+1}, y) - KL(P\hat{x}, y) + \\ KL(P\hat{x}, Px^k) + KL(x^{k+1}, x^k) - KL(Px^{k+1}, Px^k). \end{aligned} \tag{7.13}$$

From Equation (7.13) we see that the difference $KL(\hat{x}, x^k) - KL(\hat{x}, x^{k+1})$ depends only on $P\hat{x}$, and not on $\hat{x}$ itself. Summing over the index k on both sides and "telescoping" , we find that the difference $KL(\hat{x}, x^0) - KL(\hat{x}, x^*)$ also depends only on $P\hat{x}$, and not on $\hat{x}$ itself. It follows that $\hat{x} = x^*$ is the minimizer of $f(x)$ for which $KL(\hat{x}, x^0)$ is minimized. If $y = Px$ has nonnegative solutions, and the entries of x^0 are all equal to one, then x^* maximizes the Shannon entropy over all nonnegative solutions of $y = Px$.

With $f(x) = KL(Px, y)$, we have $D_f(x, z) = KL(Px, Pz)$. Therefore, we obtain the next iterate x^k by minimizing $G_k(x)$ given by

$$KL(q(x), r(x^{k-1})) = f(x) + KL(x, x^{k-1}) - D_f(x, x^{k-1}). \tag{7.14}$$

This shows that the SMART is yet another example of the "remedy" used to obtain PMAB algorithms with iterates that can be simply calculated.

The following theorem summarizes the situation with regard to the SMART [17, 18, 19].

Theorem 7.1. *In the consistent case, in which the system $y = Px$ has nonnegative solutions, the sequence of iterates of SMART converges to the unique nonnegative solution of $y = Px$ for which the distance $KL(x, x^0)$ is minimized. In the inconsistent case it converges to the unique nonnegative minimizer of the distance $KL(Px, y)$ for which $KL(x, x^0)$ is minimized. In the inconsistent case, if P and every matrix derived from P by deleting columns has full rank, then there is a unique nonnegative minimizer of $KL(Px, y)$ and at most $I - 1$ of its entries are nonzero.*

7.4. The EMML Algorithm as AM.

Now we want to minimize $f(x) = KL(y, Px)$. The iterative step of the EMML algorithm is obtained by minimizing

$$G_k(x) = KL(r(x^{k-1}), q(x)) \tag{7.15}$$

to get x^k. We have the following helpful *Pythagorean identities*:

$$KL(r(x), q(z)) = KL(r(z), q(z)) + KL(r(x), r(z)); \tag{7.16}$$

and

$$KL(r(x), q(z)) = KL(r(x), q(Mx)) + KL(Mx, z). \tag{7.17}$$

From the Pythagorean identities we have

$$KL(y, Px^k) - KL(y, Px^{k+1}) =$$

$$KL(r(x^k), r(x^{k+1})) + KL(x^{k+1}, x^k), \tag{7.18}$$

so that

$$KL(y, Px^k) - KL(y, Px^{k+1}) \geq KL(x^{k+1}, x^k). \tag{7.19}$$

The inequality in (7.19) is called the *First Monotonicity Property* in [55]. We also have $G_k(x)$ given by

$$KL(r(x), q(x)) + KL(r(x^{k-1}, r(x)) = f(x) + d(x, x^{k-1}), \tag{7.20}$$

so that

$$G_k(x) = f(x) + g_k(x), \tag{7.21}$$

with

$$d(x, x^{k-1}) = g_k(x) = KL(r(x^{k-1}), q(x)) - KL(r(x), q(x)). \tag{7.22}$$

Therefore, the EMML algorithm is an AF algorithm, so that $\{f(x^k)\}$ is decreasing. The EMML algorithm appears not to be a member of the SUMMA subclass; however, as we shall see shortly, it is a member of the SUMMA2 subclass.

Lemma 7.2. *For $\{x^k\}$ given by Equation (7.3), the sequence $\{KL(y, Px^k)\}$ is decreasing and the sequences $\{KL(x^{k+1}, x^k)\}$ and $\{KL(r(x^k), r(x^{k+1}))\}$ converge to zero.*

Lemma 7.3. *The EMML sequence $\{x^k\}$ is bounded; for $k \geq 1$ we have*

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

Using (2.3) we obtain the following useful inequality:

$$KL(r(x), r(z)) \geq KL(Mx, Mz). \tag{7.23}$$

From

$$\begin{aligned} KL(r(x), q(x^k)) &= KL(r(x^k), q(x^k)) + KL(r(x), r(x^k)) \\ &\geq f(x^k) + KL(Mx, x^{k+1}), \end{aligned}$$

and

$$\begin{gathered} KL(r(x), q(x^k)) = KL(r(x), q(Mx)) + KL(Mx, x^k) = \\ f(x) - KL(Mx, x) + KL(Mx, x^k) \end{gathered}$$

we have

$$KL(Mx, x^k) - KL(Mx, x^{k+1}) \geq f(x^k) - f(x) + KL(Mx, x). \tag{7.24}$$

Note that we have used (7.23) here. With $h_k(x) \doteq KL(Mx, x^k)$ we get Equation (5.1), so the EMML is in the SUMMA2 class. With x^* a cluster point, we have

$$KL(Mx^*, x^k) - KL(Mx^*, x^{k+1}) \geq f(x^k) - f(x^*) \geq 0. \tag{7.25}$$

Therefore, the sequence $\{KL(Mx^*, x^k)\}$ is decreasing, and the sequence $\{f(x^k)\}$ converges to $f(x^*)$. Since the EMML is in the SUMMA2 class, we know that $f(x^*)$ is the minimum value of $f(x)$ and $Mx^* = x^*$.

Let $\hat{x}$ be a minimizer of $f(x) = KL(y, Px)$. Inserting $x = \hat{x}$ into Equation (7.24), we obtain

$$KL(\hat{x}, x^k) - KL(\hat{x}, x^{k+1}) \geq KL(y, Px^k) - KL(y, Px^{k+1}). \tag{7.26}$$

The inequality in (7.26) is called the *Second Monotonicity Property* in [55].

The following theorem summarizes the situation with regard to the EMML algorithm [17, 18, 19].

Theorem 7.4. *In the consistent case, in which the system $y = Px$ has nonnegative solutions, the sequence of EMML iterates converges to a nonnegative solution of $y = Px$. In the inconsistent case it converges to a nonnegative minimizer of the distance $KL(y, Px)$. In the inconsistent case, if P and every matrix derived from P by deleting columns has full rank, then there is a unique nonnegative minimizer of $KL(y, Px)$ and at most $I - 1$ of its entries are nonzero.*

In contrast to the SMART, we have been unable to characterize the limit in terms of the starting vector x^0.

7.5. **Imposing Constraints.** In [71] we discussed certain situations in emission tomographic imaging in which it was helpful to impose reasonable constraints on the individual pixel values of the reconstructed image. In [71, 31] we presented modified versions of the SMART and EMML that employed Fermi–Dirac entropy to incorporate upper and lower bounds on these pixel values. Suppose that, for each j, we have $0 \leq a_j < b_j$ and we want to minimize $KL(Px, y)$ over all x in $X_{ab} = \{x | a_j \leq x_j \leq b_j, j = 1, ..., J\}$. In the version presented in [31] we take $s_j = 1$ for each j and

$$(7.27)\; g(x) = \sum_{j=1}^{J} \left((x_j - a_j) \log(x_j - a_j) + (b_j - x_j) \log(b_j - x_j) \right).$$

Then

$$(7.28)\quad D_g(x, z) = \sum_{j=1}^{J} \left(KL(x_j - a_j, z_j - a_j) + KL(b_j - x_j, b_j - z_j) \right).$$

It was shown there that $D_g(x, z) \geq D_f(x, z) = KL(Px, Pz)$. At the kth step of the iteration we minimize

$$(7.29)\qquad KL(Px, y) + D_g(x, x^{k-1}) - KL(Px, Px^{k-1})$$

to get x^k. If $y = Px$ has solutions satisfying the constraints, then the sequence $\{x^k\}$ converges to such a solution.

7.6. **Auxiliary Functions For Regularization.** We know from Theorems 7.1 and 7.4 that when $J > I$ and there is no nonnegative solution of $y = Px$ the limits of the SMART and EMML iterative sequences will have at least $J - I + 1$ zero values. If the x represents a vectorized reconstructed image, these zero values appear to be randomly placed in the image, making it of little value. To avoid this behavior regularization is used. By selecting the regularizing functions carefully we can still get closed-form solutions for the iterates [17]. For regularized SMART we minimize

$$(7.30)\qquad KL(Px, y) + \epsilon KL(x, p),$$

where ϵ is a small positive quantity and p is a chosen positive vector, perhaps incorporating prior information about the desired solution. To get the iterate x^k we minimize

$$KL(q(x), r(x^{k-1})) + \epsilon KL(x, p). \tag{7.31}$$

To regularize the EMML algorithm we minimize

$$KL(y, Px) + \epsilon KL(p, x). \tag{7.32}$$

To get the iterate x^k we minimize

$$KL(r(x^{k-1}), q(x)) + \epsilon KL(p, x). \tag{7.33}$$

8. Generalized Projections and Acceleration

It is well known that both the SMART and the EMML algorithm can be slow to converge. In this section we consider the use of generalized projections [4, 5, 6], row-action algorithms [39] and relaxation to accelerate convergence [20].

If, for fixed $x \geq 0$, we try to minimize $KL(x, z)$ over $z \geq 0$ with $Pz_i = y_i$, we find that we cannot obtain the desired z in closed form. However, if we minimize $\sum_{j=1}^{J} P_{i,j} KL(x_j, z_j)$, subject to $Pz_i = y_i$, we find that the desired z is $z_j = x_j y_i / Px_i$. This z is a generalized projection of x and we denote it by $z = Q_i x$. The SMART iterative step,

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

can be written as

$$x_j^k = \prod_{i=1}^{I} \left(Q_i x_j^{k-1}\right)^{P_{i,j}},$$

so that x_j^k is a weighted geometric mean of the generalized projections $Q_i x_j^{k-1}$. The EMML iterative step,

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

can be written as

$$x_j^k = \sum_{i=1}^{I} P_{i,j} (Q_i x_j^{k-1}),$$

so that x_j^k is a weighted arithmetic mean of the same generalized projections. The *multiplicative algebraic reconstruction technique* (MART) [60], with the iterative step

$$x_j^k = x_j^{k-1} \left(\frac{y_i}{(Px^{k-1})_i}\right)^{P_{i,j}}, \tag{8.1}$$

for $i = k(\operatorname{mod} I) + 1$, can be written as

$$x_j^k = \left(x_j^{k-1}\right)^{1-P_{i,j}} \left(Q_i x_j^{k-1}\right)^{P_{i,j}},$$

which says that x_j^k is a weighted geometric mean of the current x_j^{k-1} and $Q_i x_j^{k-1}$. This suggests an iterative algorithm that we have called the EMART [35], with the iterative step

$$x_j^k = (1 - P_{i,j})x_j^{k-1} + P_{i,j}(Q_i x_j^{k-1}), \tag{8.2}$$

the weighted arithmetic mean of the current x_j^{k-1} and $Q_i x_j^{k-1}$. When there are nonnegative solutions of $y = Px$ the MART converges to the same solution as the SMART. The EMART also converges to a nonnegative solution of $y = Px$, but nothing further is known.

Let $C_i \subseteq \mathbb{R}^J, i = 1, ..., I$ be nonempty closed convex sets with nonempty intersection C. The *convex feasibility problem* (CFP) [46] is to find a member of C. The generalized projections Q_i used here are defined in terms of KL distances that vary slightly with the index i. This suggests that such *multi-projection* algorithms may be used to solve the more general CFP. This idea was investigated in [24, 25, 27].

9. Block-Iterative Algorithms

More general block-iterative algorithms extending SMART and the EMML method were presented in [51, 44, 61, 11, 20, 21]. We consider these briefly in this section. Block-iterative variants of SMART and EMML will not converge to a single vector when $y = Px$ has no nonnegative solution. Therefore, when discussing convergence of block-iterative algorithms, we shall assume that $y = Px$ has nonnegative solutons. A survey of block-iterative methods is found in [31].

9.1. **Block-Iterative SMART.**

For block-iterative algorithms to solve $y = Px$ we partition the row-index set $\{i = 1, ..., I\}$ into blocks $B_1, ..., B_N$. If we do not require that the matrix P be normalized to have $s_j \doteq \sum_{i=1}^I P_{i,j} = 1$, then the SMART iterative step becomes

$$x_j^k = x_j^{k-1} \exp\left(s_j^{-1} \sum_{i=1}^I P_{i,j} \log\left(\frac{y_i}{(Px^{k-1})_i}\right)\right). \tag{9.1}$$

Obviously, all the rows of the matrix P are employed at each step of the iteration. As various authors have noted, to facilitate the use of parallel computation and to accelerate convergence it is helpful to process only some of the rows of P at each step. It may seem obvious that the appropriate block-iterative version of SMART should have the iterative step

$$x_j^k = x_j^{k-1} \exp\left(s_{n,j}^{-1} \sum_{i \in B_n} P_{i,j} \log\left(\frac{y_i}{(Px^{k-1})_i}\right)\right), \tag{9.2}$$

for $n = k (\mathrm{mod}\, N) + 1$ and $s_{n,j} = \sum_{i \in B_n} P_{i,j}$. This is not the case, however.

In [21] it was shown that block-iterative variants of the SMART with the following iterative step are convergent:

$$x_j^k = x_j^{k-1} \exp\left(\gamma_j \delta_n \sum_{i \in B_n} P_{ij} \log\left(\frac{y_i}{(Px^{k-1})_i}\right)\right), \tag{9.3}$$

where $0 < \gamma_j \delta_n s_{n,j} \leq 1$. Such iterative algorithms converge to the nonnegative solution of $y = Px$ that minimizes $\sum_{j=1}^J \gamma_j^{-1} KL(x_j, x_j^0)$, whenever such solutions exist, for any choice of blocks and any $x^0 > 0$. Since $s_{n,j}^{-1} \neq \gamma_j \delta_n$ generally, the iterative algorithm described in Equation (9.2) is not convergent.

For $\gamma_j = 1$ we must have $0 < \delta_n \leq s_{n,j}^{-1} \leq 1$, for all j, or $\delta_n \leq \min_j s_{n,j}^{-1} = 1/\max_j s_{n,j}$. With $m_n \doteq \max_j s_{n,j}$, we must have $\delta_n \leq m_n^{-1}$. The *rescaled block-iterative* SMART (RBI-SMART) [21] uses $\delta_n = m_n^{-1}$. We can write the iterative step of the RBI-SMART as

$$\log x_j^k = \log x_j^{k-1} + m_n^{-1} \sum_{i \in B_n} P_{ij} \log\left(\frac{y_i}{(Px^{k-1})_i}\right)$$

or, equivalently, as

$$\log x_j^k = (1 - s_{n,j} m_n^{-1}) \log x_j^{k-1} + m_n^{-1} \sum_{i \in B_n} P_{i,j} \log Q_i(x^{k-1})_j, \tag{9.4}$$

where, once again, we write $Q_i(x)_j = x_j \frac{y_i}{(Px)_i}$. Therefore, the RBI-SMART iterate is a weighted geometric mean of the current x^{k-1} and the generalized projections $Q_i(x^{k-1})$ for $i \in B_n$ [21, 29]. When each block consists of a single index we get the MART, so we could also call the RBI-SMART a block-iterative MART or block-MART [46] algorithm.

9.2. **Block-Iterative EMML.** The EMML has attracted more attention within the medical imaging community than has the SMART. Researchers in that field have noticed its slow convergence and have experimented with various means of acceleration. In [62, 63] the authors introduced the *ordered-subset* EM (OSEM) algorithm and observed that it often led to usable reconstructed images much faster than did the EMML.

Again, without assuming that $s_j = 1$, the EMML iterative step becomes

$$x_j^k = x_j^{k-1} s_j^{-1} \sum_{i=1}^I P_{ij} \left(\frac{y_i}{(Px^{k-1})_i}\right). \tag{9.5}$$

The OSEM is a block-iterative variant of the EMML. It uses an obvious modification of the EMML iteration and has the iterative step

$$x_j^k = x_j^{k-1} s_{n,j}^{-1} \sum_{i \in B_n} P_{ij} \left(\frac{y_i}{(Px^{k-1})_i}\right), \tag{9.6}$$

for $n = k(\bmod N) + 1$. But the OSEM is not the correct block-iterative variant of EMML; it fails to converge in most cases.

In [29] it was shown that the block-iterative algorithm with the iterative step

$$x_j^k = (1 - \gamma_j \delta_n s_{n,j}) x_j^{k-1} + x_j^{k-1} \gamma_j \delta_n \sum_{i \in B_n} P_{i,j} \left(\frac{y_i}{(Px^{k-1})_i} \right) \tag{9.7}$$

converges to a nonnegative solution of $y = Px$ for any $x^0 > 0$ and any choice of blocks, provided that $0 < \gamma_j \delta_n s_{n,j} \leq 1$. In those rare cases in which $\gamma_j \delta_n s_{n,j} = 1$ Equation (9.7) reduces to Equation (9.6). As in the RBI-SMART case, we take $\gamma_j = 1$ and $\delta_n = m_n^{-1}$ to get the RBI-EMML algorithm. The iterative step of the RBI-EMML algorithm is then

$$x_j^k = (1 - s_{n,j} m_n^{-1}) x_j^{k-1} + m_n^{-1} \sum_{i \in B_n} x_j^{k-1} P_{i,j} \left(\frac{y_i}{(Px^{k-1})_i} \right), \tag{9.8}$$

which we can write as

$$x_j^k = (1 - s_{n,j} m_n^{-1}) x_j^{k-1} + m_n^{-1} \sum_{i \in B_n} P_{i,j} Q_i(x^{k-1})_j. \tag{9.9}$$

This tells us that the RBI-EMML iterate is a weighted arithmetic mean of the current x^{k-1} and the generalized projections $Q_i(x^{k-1})$, for $i \in B_n$. As with RBI-SMART, the RBI-EMML converges, for any x^0 and any choices of blocks, to a nonnegative solution of $y = Px$. However, in contrast to RBI-SMART, we have no characterization of the particular solution to which it converges nor how that solution may vary with x^0 and the choice of blocks.

9.3. **Why Are Block-Iterative Methods Faster?** We have made the claim, and experience has shown, that in the consistent case block-iterative methods can converge significantly faster than their simultaneous relatives. We investigate this claim a bit more theoretically now. The arguments given here are not completely rigorous, but will give some idea of the source of the acceleration. Our goal is to get orders-of-magnitude estimates, not precise values. We begin by comparing the simultaneous Landweber algorithm with the sequential ART algorithm for solving the general system of linear equations $Ax = b$. Then we compare the simultaneous SMART with the sequential MART for solving the nonnegative system $Px = y$.

9.3.1. *The Landweber and Cimmino Algorithms.* Let $Ax = b$ be a consistent system of I linear equations in J unknowns, with $\sum_{j=1}^J A_{i,j}^2 = 1$, for each $i = 1, ..., I$. The iterative step of the Landweber algorithm is

$$x^{k+1} = x^k + \gamma A^T (b - Ax^k), \tag{9.10}$$

where $0 < \gamma < \frac{2}{L}$ for $L = \rho(A^T A)$, the largest eigenvalue of the matrix $A^T A$. The trace of AA^T is I, so $1 \leq L \leq I$. The choice of $\gamma = \frac{1}{I}$ is acceptable.

Simple calculations show that, for any z with $Az = b$,

$$\|z - x^k\|^2 - \|z - x^{k+1}\|^2 \geq (2\gamma - L\gamma^2) \|b - Ax^k\|^2. \tag{9.11}$$

With the choice of $\gamma = \frac{1}{I}$ we get Cimmino's algorithm:

$$x^{k+1} = x^k + \frac{1}{I} A^T (b - Ax^k), \tag{9.12}$$

and

$$\|z - x^k\|^2 - \|z - x^{k+1}\|^2 \geq (2/I - L/I^2)\|b - Ax^k\|^2. \tag{9.13}$$

The improvement we obtain in Equation (9.11) will depend L, and the choice of γ.

If we know L, which is probably not the case, especially for large systems, we may select $\gamma = \frac{1}{I}$, just to be safe; this is Cimmino's choice. If we have a better upper bound for L than just I, then we can use it in the choice of γ. For example, it was shown in [32] that, whenever the rows of A are normalized to length one, L cannot be larger than the maximum number of nonzero entries in any column of A. This is useful in the case of sparse A. In transmission tomography there are typically about $\sqrt{I}$ nonzero entries in a column, so the estimate $L \leq \sqrt{I}$ is usually acceptable. If $L = 1$ and we choose $\gamma = 1$, then Equation (9.11) becomes

$$\|z - x^k\|^2 - \|z - x^{k+1}\|^2 \geq \|b - Ax^k\|^2. \tag{9.14}$$

However, if L is closer to I than to 1 the choice of $\gamma = \frac{1}{I}$ will give us something more like

$$\|z - x^k\|^2 - \|z - x^{k+1}\|^2 \geq \frac{1}{I}\|b - Ax^k\|^2. \tag{9.15}$$

9.3.2. *The ART.* When the rows of A are normalized to have length one, the iterative step of the ART is

$$x_j^k = x_j^{k-1} + A_{i,j}(b_i - (Ax^{k-1})_i), \tag{9.16}$$

where $i = k(\operatorname{mod} I) + 1$. We consider the improvement we obtain after one pass through all the data. For any z with $Az = b$ we have

$$\|z - x^0\|^2 - \|z - x^I\|^2 = \sum_{i=1}^{I}(b_i - (Ax^{i-1})_i)^2. \tag{9.17}$$

This is, very roughly, about I times the improvement in Equation (9.15).

9.3.3. *The SMART.* For SMART we assume that $s_j = \sum_{i=1}^{I} P_{i,j} = 1$, for each j. Then, with $y = Pz$, Equation (7.13) tells us that

$$KL(z, x^k) - KL(z, x^{k+1}) \approx KL(Px^{k+1}, y). \tag{9.18}$$

9.3.4. *The MART.* With $m_i = \max\{P_{i,j} | j = 1, ..., J\}$, and $y = Pz$ we have

$$KL(z, x^0) - KL(z, x^1) \approx m_1^{-1} KL(y_1, (Px^0)_1). \tag{9.19}$$

Since $s_j = 1$, we might estimate $m_1 \approx \frac{1}{I}$. Therefore, after one pass through all the data, we have

$$KL(z, x^0) - KL(z, x^I) \approx I \, KL(y, Px^{i-1}), \tag{9.20}$$

for some representative i. The point is that the improvement we may expect after one pass through the data may well be a factor of I larger than that obtained by one SMART iteration. Of course, if the entries of P are not more or less uniformly distributed, the m_i may well be greater than $\frac{1}{I}$ and the improvement after one pass through the data may well be somewhat less than before. In the sparse case, in which there are, say, only $\sqrt{I}$ nonnegative entries in any column, the m_i will be more like $\frac{1}{\sqrt{I}}$ and the improvement will be only a factor of $\sqrt{I}$ better than SMART. Since, in many applications, I is in the thousands, even this reduced improvement is significant.

10. Probabilistic Mixture Problems

When $s_j = 1$ for all j and $x_+ \doteq \sum_{j=1}^J x_j = 1$ we can view Px as a probabilistic mixture of the columns of the matrix P, each of which is a probability vector, with the entries of x the unknown mixing proportions to be determined. In *list-mode* positron-emission tomography we sometimes encounter mixtures of probability-density functions (pdf), not of finite probability vectors [26]. A modification of the EMML algorithm, called the Mix-EM algorithm, can be used to solve this problem. In keeping with the conventions in this area we adopt somewhat different notation.

10.1. **Probabilistic-Mixture Models.**

Let X be a random vector governed by a pdf or discrete probability $f(x)$. The pdf $f(x)$ is said to be a probabilistic mixture (PM) if $f(x)$ has the form

$$f(x) = \sum_{j=1}^J \theta_j f_j(x), \tag{10.1}$$

where the $f_j(x)$ are known probability-density functions (pdf) or finite or infinite discrete probabilities and the entries of $\theta = (\theta_1, ..., \theta_J)^T$ are to be determined, subject to $\sum_{j=1}^J \theta_j = 1$. We have finitely many realizations of X, denoted $x_1, ..., x_N$, from which we must estimate the mixing proportions θ_j. The estimate is obtained by maximizing the likelihood function

$$L(\theta) = \prod_{n=1}^N f(x_n) = \prod_{n=1}^N \left(\sum_{j=1}^J \theta_j f_j(x_n) \right).$$

In a more general formulation of probabilistic mixture the pdf $f_j(x)$ involve parameters to be determined as well. For example, we may wish to model $f(x)$ as a probabilistic mixture of a small number of normal pdf whose means and variances are unknown, or a small number of Poisson probabilities with unknown means. If the f_j themselves involve unknown parameters, we have a choice: we can take J large, but expect only a few of the θ_j to be non-zero; or we can estimate J, the non-zero θ_j, and the parameters associated with those f_j for which θ_j is non-zero. In most applications of PM models the pdf $f(x)$ is suspected of being a superposition of a relatively small number of components $f_j(x)$ and the goal is to determine the relative sizes of the θ_j and, in particular, which θ_j are non-zero [14, 16]. Such models have uses in a wide variety of applications, including sonar, radar, astronomy, spectral analysis, analytic chemistry, and many others.

10.2. **The Mix-EM Algorithm.** When the f_j are probability-density functions the values $f_j(x_n)$ can be any nonnegative values. In such cases we cannot apply the EMML algorithm directly, just by replacing $P_{i,j}$ with $f_j(x_n)$. We require a modification of the EMML algorithm that we call the Mix-EM algorithm [26]. We start with $\theta_j^0 > 0$, for each j. Having found θ_j^{k-1}, for each j, we define

$$\theta_j^k = \theta_j^{k-1} \frac{1}{N} \sum_{n=1}^{N} \left(f_j(x_n) \frac{1}{(P\theta^{k-1})_n} \right), \tag{10.2}$$

where P is the matrix with entries $P_{n,j} \doteq f_j(x_n)$ and

$$(P\theta^{k-1})_n = \sum_{j=1}^{J} P_{n,j} \theta_j^{k-1}. \tag{10.3}$$

The sequence $\{\theta^k\}$ converges to a vector of maximum-likelihood values of the θ_j [26].

11. Fixed-Point Algorithms

Suppose that $T : X \to X$ is an operator on the set X. We say that z is a *fixed point* for T if $Tz = z$. A variety of problems can be solved using fixed-point iteration, in which $x^k = Tx^{k-1}$ and T is selected so that the solutions of the problem coincide with the fixed points of T. We have seen this already in this paper, in our discussion of the FBS methods, the SMART and the EMML algorithm. Most of the theory of fixed-point iteration is developed within the context of X a Hilbert space [5, 28, 38], although, as we have seen, this approach can be applied more generally [10, 57, 17, 19, 46, 6].

Fixed-point theory seems to play little role in AF methods, although a few things can be said. Consider a SUMMA iteration algorithm in which we minimize $f(x) + g_k(x)$ to get x^k. Suppose that $x^k = x^{k-1}$ for some k. Then we must have $f(x^k) = \beta = \min_{x \in C} f(x)$. In PMAB iteration we can define an operator T by setting Tz equal to the minimizer of $f(x) + D_h(x, z)$. If $Tz = z$, then $f(z) = \beta$ so z is a solution to our problem.

12. Some Questions

In this section we survey a few open questions related to the topics discussed in this article.

12.1. Limit Cycles for MART and EMART. If the M by N system of linear equations $Ax = b$ has no solution then the *algebraic reconstruction technique* (ART) iteration [60], in which $x^k = P_M P_{M-1} \cdots P_2 P_1 x^{k-1}$, cannot converge to a single vector, where P_m denotes the orthogonal projection onto the hyperplane $H_m = \{x | (Ax)_m = b_m\}$. Tanabe shows in [76] that we get subsequential convergence to a limit cycle of (usually) M distinct vectors $z^1, ..., z^M$, with $P_1 z^M = z^1$ and $P_m z^{m-1} = z^m$, for $m = 2, 3, ..., M$. When the nonnegative system of linear equations $y = Px$ has no nonnegative solution neither the MART nor the EMART can converge to a single vector. In practice we always see subsequential convergence to limit cycles, but as yet no proof of their existence has emerged.

12.2. The Goldstein–Osher Problem. The following question arises from some assertions made in [59]. Suppose that x^k minimizes

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

and $\{x^k\}$ converges to x^*. We know that x^* minimizes $f(x)$ over all x in C, the domain of the function f. Let S be the set of all such minimizers. Does x^* also minimize $h(z)$ over all z in S? In general, the answer is no; D_h does not determine h uniquely. What if $h(x) = D_h(x, x^0)$; that is, what if $h(x^0) = 0$ and $\nabla h(x^0) = 0$? There are several examples, using both Euclidean and Kullback-Leibler distances, in which the answer is yes.

12.3. Characterizing the EMML and EMART Solutions. When the nonnegative system of linear equations $y = Px$ has nonnegative solutions then SMART and MART converge to the nonnegative solution that minimizes $KL(x, x^0)$, where x^0 is the positive starting vector for the iteration. It has been shown that both EMML and EMART converge to nonnegative solutions in this case, but no similar characterization of the limit is known.

References

[1] S. Ahn, J. Fessler, D. Blatt and A. Hero *Convergent incremental optimization transfer algorithms: application to tomography*, IEEE Trans. Med. Imaging **25(3)** (2006), 283–296.

[2] A. Auslender and M. Teboulle *Interior gradient and proximal methods for convex and conic optimization*, SIAM J. Optimization **16(3)** (2006), 697–725.

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

[4] H. Bauschke and J. Borwein *On the convergence of von Neumann's alternating projection algorithm for two sets*, Set-Valued Analysis **1** (1993), 185–212.

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

[6] H. Bauschke and J. Borwein *Legendre functions and the method of random Bregman projections*, J. Convex Anal. **4** (1997), 27–67.

[7] H. Bauschke and J. Borwein *Joint and separate convexity of the Bregman distance*, in [12] (2001), 23–36.

[8] H. Bauschke, P. Combettes and D. Noll *Joint minimization with alternating Bregman proximity operators*, Pacific J. Optim. **2** (2006), 401–424.

[9] H. Bauschke and P. Combettes *The Baillon-Haddad Theorem revisited*, J. Convex Anal. **17** (2010), 781–787.

[10] L. M. Bregman *The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming*, USSR Comput. Math. and Math. Phys. **7** (1967), 200–217.

[11] J. Browne and A. DePierro *A row-action alternative to the EM algorithm for maximizing likelihoods in emission tomography*, IEEE Trans. Med. Imag. **15** (1996), 687–699.

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

[13] D. Butnariu, C. Byrne and Y. Censor *Redundant axioms in the definition of Bregman functions*, J. Convex Anal. **10** (2003), 245–254.

[14] C. Byrne, B.M. Levine and J.C. Dainty *Stable estimation of the probability density function of intensity from photon frequency counts*, JOSA Communications **1(11)** (1984), 1132–1135.

[15] C. Byrne and M. Fiddy *Estimation of continuous object distributions from Fourier magnitude measurements*, JOSA A **4** (1987), 412–417.

[16] C. Byrne, D. Haughton and T. Jiang *High-resolution inversion of the discrete Poisson and binomial transformations*, Inverse Problems **9** (1993), 39–56.

[17] C. Byrne *Iterative image reconstruction algorithms based on cross-entropy minimization*, IEEE Trans. Image Proc. **IP-2** (1993), 96–103.

[18] C. Byrne *Erratum and addendum to 'Iterative image reconstruction algorithms based on cross-entropy minimization'*, IEEE Trans. Image Proc. **IP-4** (1995), 225–226.

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

[20] C. Byrne (1996) *Block-iterative methods for image reconstruction from projections*, IEEE Trans. Image Proc. **IP-5** (1996), 792–794.

[21] C. Byrne *Accelerating the EMML algorithm and related iterative algorithms by rescaled block-iterative (RBI) methods*, IEEE Trans. Image Proc. **IP-7** (1998), 100–109.

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

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

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

[25] C. Byrne *Bregman-Legendre multidistance projection algorithms for convex feasibility and optimization*, in [12] (2001), 87–99.

[26] C. Byrne *Likelihood maximization for list-mode emission tomographic image reconstruction*, IEEE Trans. Med. Imag. **20(10)** (2001), 1084–1092.

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

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

[29] C. Byrne *Choosing parameters in block-iterative or ordered-subset reconstruction algorithms*, IEEE Trans. Image Proc. **14 (3)** (2005), 321–327.

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

[31] C. Byrne *Block-iterative algorithms*, International Trans. Operations Research **16(4)** (2009), 1–37.

[32] C. Byrne *Bounds on the largest singular value of a matrix and the convergence of simultaneous and block-iterative algorithms for sparse linear systems*, International Trans. Operations Research **16(4)** (2009), 465–479.

[33] C. Byrne *Alternating minimization as sequential unconstrained minimization: a survey*, J. Opt. Th. and Appl., electronic **154(3)**, DOI 10.1007/s1090134-2, (2012), and hardcopy **156(3)**, February, 2013, 554–566.

[34] C. Byrne *An elementary proof of convergence of the forward-backward splitting algorithm*, J. Nonlinear and Convex Anal. **15(4)** (2014), 681–691.

[35] C. Byrne *Iterative Optimization in Inverse Problems*. Boca Raton, FL: CRC Press, 2014.

[36] C. Byrne *Signal Processing: A Mathematical Approach: 2nd ed.*. Boca Raton, FL: CRC Press, 2014.

[37] C. Byrne *On a generalized Baillon–Haddad Theorem for convex functions on Hilbert space*, J. Convex Anal. **22(4)** (2015), 963–967.

[38] A. Cegielski *Iterative Methods for Fixed Point Problems in Hilbert Space*. Heidelberg: Springer Lecture Notes in Mathematics 2057, 2012.

[39] Y. Censor *Row-action methods for huge and sparse systems and their applications*, SIAM Review **23** (1981), 444–464.

[40] Y. Censor and T. Elfving *A multi-projection algorithm using Bregman projections in a product space*, Numerical Algorithms **8** (1994), 221–239.

[41] Y. Censor and S. Reich *Iterations of paracontractions and firmly nonexpansive operators with applications to feasibility and optimization*, Optimization **37** (1996), 323–339.

[42] Y. Censor, T. Bortfeld, B. Martin and A. Trofimov *A unified approach for inversion problems in intensity-modulated radiation therapy*, Phys. Med. and Biol. **51** (2006), 2353–2365.

[43] Y. Censor, T. Elfving, N. Kopf and T. Bortfeld *The multiple-sets split feasibility problem and its application for inverse problems*, Inverse Problems **21** (2005), 2071–2084.

[44] Y. Censor and J. Segman *On block-iterative maximization*, J. Inform. and Optim. Sci. **8** (1987), 275–291.

[45] Y. Censor and S.A. Zenios *Proximal minimization algorithm with D-functions*, J. Optim. Th. and Appl. **73(3)** (1992), 451–464.

[46] Y. Censor and S.A. Zenios *Parallel Optimization: Theory, Algorithms and Applications*, New York: Oxford University Press, 1997.

[47] W. Cheney and A. Goldstein *Proximity maps for convex sets*, Proc. Amer. Math. Soc. **10** (1959), 448–450.

[48] E. Chi, H. Zhou and K. Lange *Distance majorization and its applications*, Math. Program. **146 (1-2)** (2014), 409–436.

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

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

[51] J. Darroch and D. Ratcliff *Generalized iterative scaling for log-linear models*, Annals of Math. Stat. **43** (1972), 1470–1480.

[52] A.P. Dempster, N.M. Laird and D.B. Rubin *Maximum likelihood from incomplete data via the EM algorithm*, J. Roy. Stat. Soc., Series B **37** (1977), 1–38.

[53] A. De Pierro *A modified expectation maximization algorithm for penalized likelihood estimation in emission tomography*, IEEE Trans. Med. Imag. **14** (1995), 132–137.

[54] A. De Pierro and M. Yamaguchi *Fast EM-like methods for maximum 'a posteriori' estimates in emission tomography*, IEEE Trans. Med. Imag. **20(4)**, 280–288.

[55] P. Eggermont and V. LaRiccia *On EM-like algorithms for minimum distance estimation*, http://www.udel.edu/FREC/eggermont/Preprints/emlike.pdf (1998).

[56] P. Eggermont and V. LaRiccia *Maximum Penalized Likelihood Estimation.* New York: Springer 2001.

[57] L. Elsner, L. Koltracht and M. Neumann *Convergence of sequential and asynchronous nonlinear paracontractions*, Numerische Mathematik **62** (1992), 305–319.

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

[59] T. Goldstein and S. Osher *The split Bregman algorithm for L^1 regularized problems*, SIAM J. Imaging Sciences **2(2)** (2009), 323–343.

[60] R. Gordon, R. Bender and G.T. Herman *Algebraic reconstruction techniques (ART) for three-dimensional electron microscopy and x-ray photography*, J. Theoret. Biol. **29** (1970), 471–481.

[61] S. Holte, P. Schmidlin, A. Linden, G. Rosenqvist and L. Eriksson *Iterative image reconstruction for positron emission tomography: a study of convergence and quantitation problems*, IEEE Trans. Nucl. Sci. **37** (1990), 629–635.

[62] M. Hudson, B. Hutton and R. Larkin *Accelerated EM reconstruction using ordered subsets*, J. Nucl. Med. **33** (1992), 960.

[63] M. Hudson and R. Larkin *Accelerated image reconstruction using ordered subsets of projection data*, IEEE Trans. Med. Imag. **13** (1994), 601–609.

[64] S. Kullback and R. Leibler *On information and sufficiency*, Annals of Math. Stat. **22** (1951), 79–86.

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

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

[67] G. McLachlan and T. Krishnan *The EM Algorithm and Extensions.* New York: John Wiley and Sons, Inc. 1997.

[68] J.-J. Moreau *Fonctions convexes duales et points proximaux dans un espace hilbertien*, C.R. Acad. Sci. Paris Sér. A Math. **255** (1962), 2897–2899.

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

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

[71] M. Narayanan, C. Byrne and M. King *An interior point iterative maximum-likelihood reconstruction algorithm incorporating upper and lower bounds, with application to SPECT transmission imaging*, IEEE Trans. Med. Imag. **TMI-20(4)** (2001), 342–353.

[72] S. Penfold, R. Zalas, M. Casiraghi, M. Brooke, Y. Censor and R. Schulte *Sparsity constrained split feasibility for dose-volume constraints in inverse planning of intensity-modulated photon or proton therapy*, Phys. Med. Biol. **62** (2017), 3599-3618. DOI:10.1088/1361-6560/aa602b.

[73] A. Rockmore and A. Macovski *A maximum likelihood approach to emission image reconstruction from projections*, IEEE Trans. Nucl. Sci. **NS-23** (1976), 1428–1432.

[74] P. Schmidlin *Iterative separation of sections in tomographic scintigrams*, Nuklearmedizin **11** (1972), 1–16.

[75] L. Shepp and Y. Vardi *Maximum likelihood reconstruction for emission tomography*, IEEE Trans. Med. Imag. **MI-1** (1982), 113–122.

[76] K. Tanabe *Projection method for solving a singular system of linear equations and its applications*, Numer. Math. **17** (1971), 203–214.

[77] Y. Vardi, L. Shepp and L. Kaufman *A statistical model for positron emission tomography*, J. Amer. Stat. Assoc. **80** (1985), 8–20.

(C. Byrne) Department of Mathematical Sciences, University of Massachusetts Lowell, Lowell, MA, USA
E-mail address: `Charles_Byrne@uml.edu`