Alternating Minimization and Alternating Projection Algorithms: A Tutorial

Charles L. Byrne
Department of Mathematical Sciences
University of Massachusetts Lowell
Lowell, MA 01854

March 27, 2011

Abstract

Alternating minimization of a function of two vectors variables provides a useful framework for the derivation of iterative optimization algorithms. The main reference for alternating minimization is the paper [32] of Csiszár and Tusnády. We use their three-point property and four-point property to provide a somewhat simpler proof of convergence for their alternating minimization algorithm. Examples include alternating orthogonal projection between closed convex sets in R^N and alternating entropic projection between closed convex sets in R^N_+; The SMART and EMML iterative algorithms are special cases of the latter. Extensions of these notions to alternating orthogonal and generalized projection onto convex sets, the convex feasibility problem and the split feasibility problem are also considered.

1 Alternating Minimization

Alternating minimization of a function of two vectors variables provides a useful framework for the derivation of iterative optimization algorithms. The main reference for alternating minimization (alt min) is the paper [32] of Csiszár and Tusnády. As the authors of [55] remark, the geometric argument in [32] is "deep, though hard to follow". In this section we use three-point property and four-point property of [32] to provide a somewhat simpler proof of convergence for their alternating minimization algorithm.

1.1 The Basic Alt Min Framework

Suppose that P and Q are arbitrary 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,

$$d = \inf_{p \in P,\, q \in Q} \Theta(p, q) < +\infty.$$

We assume that $d > -\infty$; in most applications, the function $\Theta(p, q)$ is non-negative, so this additional assumption is unnecessary. We do not always assume there are $\hat{p} \in P$ and $\hat{q} \in Q$ such that

$$\Theta(\hat{p}, \hat{q}) = d;$$

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

$$\Theta(p^n, q^n) \to d.$$

The *alternating minimization algorithm* 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 a later section we consider the special case of alternating cross-entropy minimization. In that case, the vectors p and q are non-negative, 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 that particular application 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 d, since we have

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

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

We know two things:

$$\Theta(p^n, q^{n-1}) - \Theta(p^n, q^n) \geq 0, \tag{1.1}$$

and

$$\Theta(p^n, q^n) - \Theta(p^{n+1}, q^n) \geq 0. \tag{1.2}$$

We need to make these inequalities more precise.

1.2 The Three- and Four-Point Properties

Suppose that $\Delta : P \times P \to R$ is a non-negative function with $\Delta(p, p) = 0$ for all $p \in P$.

Definition 1.1 *Say that $p \in P$ has the* three-point property *if, for $n = 1, 2, ...$, we have*

$$\Theta(p, q^n) \geq \Theta(p^{n+1}, q^n) + \Delta(p, p^{n+1}), \tag{1.3}$$

When the three-point property holds, the inequality in (1.2) is strengthened to

$$\Theta(p^n, q^n) - \Theta(p^{n+1}, q^n) \geq \Delta(p^n, p^{n+1}). \tag{1.4}$$

Suppose that p^n has the three-point property for all n. Then, from the inequalities

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

we see not only that the sequence $\{\Theta(p^n, q^n)\}$ is decreasing, but that the sequence $\{\Delta(p^n, p^{n+1})\}$ converges to zero. To prove the main theorem we need one more property.

Definition 1.2 *Say that p has the* four-point property *if, for all $q \in Q$ and all n,*

$$\Delta(p, p^n) + \Theta(p, q) \geq \Theta(p, q^n). \tag{1.5}$$

When the four-point property holds, the inequality in (1.1) is extended to

$$\Delta(p^n, p^{n-1}) \geq \Theta(p^n, q^{n-1}) - \Theta(p^n, q^n) \geq 0. \tag{1.6}$$

We shall be particularly interested in cases in which $\hat{p}$ has the three- and four-point property. Then we shall have

$$\Theta(\hat{p}, q^n) \geq \Theta(p^{n+1}, q^n) + \Delta(\hat{p}, p^{n+1}), \tag{1.7}$$

and

$$\Delta(\hat{p}, p^n) + \Theta(\hat{p}, \hat{q}) \geq \Theta(\hat{p}, q^n). \tag{1.8}$$

Combining these two inequalities, we obtain the inequality

$$\Delta(\hat{p}, p^n) - \Delta(\hat{p}, p^{n+1}) \geq \Theta(p^{n+1}, q^n) - \Theta(\hat{p}, \hat{q}) \geq 0. \tag{1.9}$$

Now we are ready for the main theorem.

1.3 The Main Theorem

We want to find sufficient conditions for the sequence $\{\Theta(p^n, q^n)\}$ to converge to d, that is, for $D = d$. Suppose that $D > d$. Then there are p' and q' such that

$$D > \Theta(p', q') \geq d.$$

If p' has the three- and four-point properties, then

$$\Theta(p', q^0) \geq \Theta(p^1, q^0) + \Delta(p', p^1),$$

and

$$\Delta(p', p^1) \geq \Theta(p', q^1).$$

Since we have assumed that we have selected q^0 so that $\Theta(p, q^0)$ is finite for all p, it follows that, if p' has the three- and four-point properties, then $\Theta(p', q^n)$ and $\Delta(p', p^n)$ are finite for all n. The main theorem is the following.

Theorem 1.1 *Suppose that p has the three-point and four-point properties, for any p for which there is a q with $\Theta(p^n, q^n) \geq \Theta(p, q)$ for all n. Then $D = d$ and so*

$$\Theta(p^n, q^n) \to d. \tag{1.10}$$

Proof: Suppose that $D > d$ and p' and q' are as above. Then we have

$$\Delta(p', p^n) - \Delta(p', p^{n+1}) \geq \Theta(p^{n+1}, q^n) - \Theta(p', q') \geq 0. \tag{1.11}$$

We know that $\Delta(p', p^{n+1})$ is finite for each n. It follows that the right side of (1.11) converges to zero, being the difference of successive terms of a decreasing, non-negative sequence. Therefore, $D = \Theta(p', q')$, which is a contradiction. So $D = d$. ∎

Corollary 1.1 *If there are $\hat{p}$ and $\hat{q}$ such that $\Theta(\hat{p}, \hat{q}) = d$, $\Theta(\hat{p}, q^0)$ is finite, and $\hat{p}$ has the three- and four-point properties, then*

$$\{\Theta(p^n, q^n)\} \to \Theta(\hat{p}, \hat{q}).$$

We know, therefore, that if the three- and four-point properties hold, then

$$\Theta(p^n, q^n) \to \Theta(\hat{p}, \hat{q}).$$

Up to now we have said nothing about convergence of the $\{p^n\}$ or $\{q^n\}$ themselves; for that, we need to assume some topology on the sets P and Q and make further assumptions.

1.4 Possible Additional Assumptions

Here are some assumptions that we may make in order to establish convergence of the sequences $\{p^n\}$ and $\{q^n\}$.

- P and Q are subsets of (possibly different) metric spaces, such as R^N;
- if $\{\Delta(p, p^n)\}$ is decreasing, then $\{p^n\}$ has a subsequence $\{p^{n_k}\}$ converging to p^* and $\{q^{n_k}\}$ converges to q^*;
- the sequence $\{\Theta(p^{n_k}, q^{n_k})\}$ converges to $\Theta(p^*, q^*)$, so $\Theta(p^*, q^*) = d$;
- the sequence $\{\Delta(p^*, p^{n_k})\}$ converges to $\Delta(p^*, p^*) = 0$;
- if $\{\Delta p^*, p^n)\}$ converges to zero, then $\{p^n\}$ converges to p^*.

Substituting p^* for $\hat{p}$, we find that the sequence $\{\Delta(p^*, p^n)\}$ is decreasing, since $d = \Theta(p^*, q^*) \leq \Theta(p^n, q^n)$ for all n. Therefore, the sequence $\{\Delta(p^*, p^n)\}$ converges to zero, since a subsequence converges to zero. We conclude then that $\{p^n\}$ converges to p^*.

2 Alternating Euclidean Distance Minimization

Let P and Q be non-empty closed convex subsets of R^N. Our objective is to minimize the function $\|p - q\|^2$ over all $p \in P$ and $q \in Q$. If $P \cap Q \neq \emptyset$, then the alternating minimization method will find a member of the intersection.

Let $\Theta(p, q) = \|p - q\|^2$ and $\Delta(p, p') = \|p - p'\|^2$. Then p^{n+1} is the orthogonal projection of q^n onto P and q^{n+1} is the orthogonal projection of p^{n+1} onto Q.

For any closed convex subset C the orthogonal projection operator P_C has the characterizing property that

$$\langle P_C x - x, c - P_C x \rangle \geq 0, \tag{2.1}$$

for all $c \in C$. From this it follows that

$$\|P_C x - P_C z\| \leq \|x - z\|. \tag{2.2}$$

for all x and z. If $f : R^N \to R$ is differentiable, and there is $x \in C$ such that $f(x) \leq f(c)$, for all $c \in C$, then

$$\langle \nabla f(x), c - x \rangle \geq 0,$$

for all $c \in C$.

2.1 The Three- and Four-Point Properties Hold

First, we show that the three-point property holds for all p. We have

$$\|p - q^n\|^2 = \|p - p^{n+1} + p^{n+1} - q^n\|^2$$

$$= \|p - p^{n+1}\|^2 + \|p^{n+1} - q^n\|^2 + 2\langle p - p^{n+1}, p^{n+1} - q^n\rangle$$

so that

$$\|p - q^n\|^2 - \|p^{n+1} - q^n\|^2 \geq \|p - p^{n+1}\|^2.$$

This is the three-point property.

Now, we show that the four-point property holds for all p. According to the authors of [32], the four-point property comes from another simple application of the inequality in (2.1), but it appears to require a bit of calculation. Our goal is to show that

$$\|p - p^n\|^2 + \|p - q\|^2 \geq \|p - q^n\|^2,$$

for all p and q, and for all n.

We have

$$\|p - p^n\|^2 = \|p - q + q - p^n\|^2 = \|p - q\|^2 + \|q - p^n\| + 2\langle p - q, q - p^n\rangle$$

$$= \|p - q\|^2 + \|q - q^n + q^n - p^n\|^2 + 2\langle p - q, q - p^n\rangle$$

$$= \|p - q\|^2 + \|q - q^n\|^2 + \|q^n - p^n\|^2 + 2\langle q - q^n, q^n - p^n\rangle + 2\langle p - q, q - p^n\rangle$$

$$\geq \|p - q\|^2 + \|q - q^n\|^2 + \|q^n - p^n\|^2 + 2\langle p - q, q - p^n\rangle.$$

Next, we use

$$\|q - q^n\|^2 = \|q - p + p - q^n\|2 = \|q - p\|^2 + \|p - q^n\|^2 + 2\langle q - p, p - q^n\rangle,$$

and

$$2\langle q - p, p - q^n\rangle + 2\langle p - q, q - p^n\rangle = -2\|p - q\|^2 + 2\langle p - q, q^n - p^n\rangle$$

to get

$$\|p - p^n\|^2 \geq \|p - q^n\|^2 + \|q^n - p^n\|^2 + 2\langle p - q, q^n - p^n\rangle.$$

Finally, we use

$$\|q^n - p^n\|^2 + 2\langle p - q, q^n - p^n\rangle = \|p - q + q^n - p^n\|^2 - \|p - q\|^2$$

to get

$$\|p - p^n\|^2 + \|p - q\|^2 \geq \|p - q^n\|^2 + \|p - q + q^n - p^n\|^2,$$

from which the four-point property follows immediately.

From the main theorem we can conclude that the sequence $\{\|p^n - q^n\|^2\}$ converges to d. It is possible for $d = 0$ without there being any $\hat{p}$ and $\hat{q}$. For example, let R^N be R^2, P be the epi-graph of the function $y = 1/x$, for $x > 0$, and Q the closed lower half-plane. If there are $\hat{p}$ and $\hat{q}$ with $\|\hat{p} - \hat{q}\|^2 = d$, then, from the three- and four-point properties, we have that the sequences $\{\|\hat{p} - p^n\|^2\}$ and $\{\|\hat{p} - q^n\|^2\}$ are decreasing, so that $\{p^n\}$ and $\{q^n\}$ are bounded. It is then easy to show that $\{p^n\} \to p^*$ and $\{q^n\} \to q^*$, with $\|p^* - q^*\|^2 = d$. This result was given by Cheney and Goldstein in 1959 [27] (see also [2]).

2.2 Approximate Methods

We have implicitly assumed, throughout this section, that the orthogonal projection onto P and Q are easily calculated. In general, this will not be the case, and approximate methods will be needed. One such approximate method is, at each step, to replace the orthogonal projection onto P or Q with orthogonal projection onto a supporting hyperplane. We shall discuss this approach in more detail when we consider the split feasibility problem.

3 A Landweber-Type Algorithm

Using the theory of the previous section, we derive an iterative algorithm for solving a system $Ax = b$ of linear equations. Such algorithms are useful when the number of equations and unknowns is large. The algorithm we obtain is a special case of Landweber's algorithm [45] (see also [16]).

3.1 The Alt Min Formulation

Let $N = IJ$ and P be the set of all I by J arrays $p = \{p_{ij}\}$ in R^N such that $b_i = \sum_{j=1}^{J} p_{ij}$ for all i. Let Q be the set of all I by J arrays in R^N of the form $q = q(x) = \{q_{ij} = A_{ij}x_j\}$ for some vector x. The subsets P and Q are non-empty, closed and convex in R^N.

First, we minimize $\|p - q(x)\|$ over all $p \in P$. Using Lagrange multipliers, we find that the optimal p must satisfy the equations

$$p_{ij} = A_{ij}x_j + \lambda_i,$$

for some constant λ_i. Summing over the index j, we find that

$$b_i = (Ax)_i + J\lambda_i,$$

for each i. Therefore,

$$\lambda_i = \frac{1}{J}(b_i - (Ax)_i),$$

so that the optimal p has the entries

$$p_{ij} = A_{ij}x_j + \frac{1}{J}(b_i - (Ax)_i);$$

we denote this p by $p(x)$. Note that

$$\|p(x) - q(x)\|^2 = \|b - Ax\|^2.$$

Now we minimize $\|p(x) - q(z)\|$ with respect to z.

Setting the z-gradient to zero, we have

$$0 = (\sum_{i=1}^{I} A_{ij}^2)(x_j - z_j) + \frac{1}{J}\sum_{i=1}^{I} A_{ij}(b_i - (Ax)_i),$$

for each j. Therefore, the optimal z has the entries

$$z_j = x'_j = x_j + \alpha_j \sum_{i=1}^{I} A_{ij}(b_i - (Ax)_i),$$

with

$$\alpha_j = (J\sum_{i=1}^{I} A_{ij}^2)^{-1}.$$

3.2 The Iterative Algorithm

Our algorithm then has the iterative step

$$x_j^{n+1} = x_j^n + \alpha_j \sum_{i=1}^{I} A_{ij}(b_i - (Ax)_i), \tag{3.1}$$

for each j. The sequence $\{x^n\}$ converges to a minimizer of the function $\|b - Ax\|$, as a consequence of our previous results. We can also establish convergence by relating this iterative method to that of Landweber.

3.3 Using The Landweber Algorithm

Let

$$\beta_j = \alpha_j^{-1} = J\sum_{i=1}^{I} A_{ij}^2.$$

Define the matrix B to have the entries

$$B_{ij} = A_{ij}\sqrt{\beta_j},$$

and

$$z_j = x_j\sqrt{\alpha_j}.$$

Then $Bz = Ax$ and the iteration in Equation (3.1) becomes

$$z_j^{n+1} = z_j^n + \sum_{i=1}^{I} B_{ij}(b_i - (Bz^n)_i), \tag{3.2}$$

for each j. The Landweber iteration for the system $Bz = b$ is

$$z_j^{n+1} = z_j^n + \gamma \sum_{i=1}^{I} B_{ij}(b_i - (Bz^n)_i).$$

It is well known (see, for example, [16]), that the Landweber iterative sequence converges to the minimizer of $\|b - Bz\|$ closest to z^0, provided that $0 < \gamma < \frac{2}{L}$, where L is the largest eigenvalue of the matrix B^TB. Since the trace of B^TB is one, it follows that the choice of $\gamma = 1$ is acceptable and that the iterative sequence generated by Equation (3.2) converges to the minimizer of $\|b - Bz\|$ closest to z^0. Therefore, our original iterative sequence $\{x^n\}$ converges to the minimizer of $\|b - Ax\|$ for which the function

$$\sum_{j=1}^{J} |\beta_j(x_j - x_j^0)|^2$$

is minimized. The Cimmino algorithm [28] is a special case of the Landweber algorithm.

3.4 The Projected Landweber Algorithm

Suppose that we want to find a minimizer of the function $\|b - Ax\|$ over $x \in C$, where C is a non-empty, closed convex subset of R^J. We can then use the *projected Landweber* theory [16], which tells us that the sequence defined by

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

converges to a solution of this problem.

The projected Landweber algorithm can be viewed as successive orthogonal projection onto three distinct closed convex sets, P, Q, and finally C. Here our goal is to find a member of the intersection of these three sets. Later we shall discuss the *convex feasibility problem*, which involves any finite number of closed convex sets. Successive orthogonal projection works in this more general case as well.

3.5 Accelerating Convergence

In our use of Landweber's algorithm we employed the trace of the matrix B^TB as our estimate of L, the spectral radius. In many applications this estimate is too high, particularly when B is sparse. Tighter upper bounds for L were presented in [18] and we can use these results to accelerate the convergence of the iteration in Equation (3.1).

For $i = 1, ..., I$ let s_i denote the number of non-zero entries in the ith row of the matrix A. For $j = 1, ..., J$ let

$$t_j = \sum_{i=1}^{I} s_i A_{ij}^2,$$

and t be the maximum of the t_j. It was shown in [18] that $L \leq t$. It follows that the iteration defined by

$$x_j^{n+1} = x_j^n + t_j^{-1} \sum_{i=1}^{I} A_{ij}(b_i - (Ax^n)_i) \tag{3.3}$$

converges to a minimizer of $\|b - Ax\|$. This iterative method is closely related to the CAV algorithm in [26].

Note that for sparse matrices s_i will be much less than J, so the increment $x_j^{n+1} - x_j^n$ in Equation (3.3) will be much larger than the corresponding one in Equation (3.1). This will tend to accelerate the convergence of the iteration.

4 Alternating Entropic Distance Minimization

Now we suppose that P and Q are closed convex subsets of R_+^N and we want to minimize a different distance, the cross-entropy or Kullback-Leibler distance between $p \in P$ and $q \in Q$.

4.1 The Kullback-Leibler Distance

For $a > 0$ and $b > 0$, the Kullback-Leibler distance, $KL(a, b)$, is defined as

$$KL(a, b) = a \log \frac{a}{b} + b - a. \tag{4.1}$$

In addition, $KL(0, 0) = 0$, $KL(a, 0) = +\infty$ and $KL(0, b) = b$. The KL distance is then extended to nonnegative vectors coordinate-wise [44]. The following lemma is easy to prove.

Lemma 4.1 *For $x \in R^J$ let $x_+ = \sum_{j=1}^{J} x_j$. Then*

$$KL(z, x) \geq KL(z_+, x_+).$$

4.2 Minimizing $KL(p, q)$

Now we suppose that P and Q are closed convex subsets of R_+^N and we want to minimize $KL(p, q)$ over $p \in P$ and $q \in Q$. We apply the alternating minimization formulation, with

$$\Theta(p, q) = KL(p, q),$$

and

$$\Delta(p, p') = KL(p, p').$$

We assume that we have been able to select a starting vector $q^0 \in Q$ such that $KL(p, q^0)$ is finite for all $p \in P$.

4.3 The Three-Point Property Holds

Let p^1 minimize $KL(p, q^0)$ over all $p \in P$. The partial derivative of $f(p) = KL(p, q^0)$ with respect to p_n is

$$\frac{\partial f}{\partial p_n} = \log \frac{p_n}{q_n^0},$$

so that

$$\sum_{n=1}^{N} (p_n - p_n^1) \log \frac{p_n^1}{q_n^0} \geq 0,$$

for all $p \in P$. Then

$$KL(p, q^0) - KL(p, p^1) = \sum_{n=1}^{N} \left(p_n \log \frac{p_n^1}{q_n^0} + q_n^0 - p_n^1 \right)$$

$$= \sum_{n=1}^{N} \left(p_n^1 \log \frac{p_n^1}{q_n^0} + q_n^0 - p_n^1 \right) + \sum_{n=1}^{N} (p_n - p_n^1) \log \frac{p_n^1}{q_n^0} \geq KL(p^1, q^0).$$

So the three-point property holds for all p.

4.4 The Four-Point Property Holds

We know that q^1 minimizes $KL(p^1, q)$ over all $q \in Q$, and that the partial derivative of $g(q) = KL(p^1, q)$ with respect to q_n is

$$\frac{\partial g}{\partial q_n} = -\frac{p_n^1}{q_n} + 1.$$

Therefore,

$$\sum_{n=1}^{N} (q_n^1 - q_n) \frac{p_n^1}{q_n^1} \geq \sum_{n=1}^{N} (q_n^1 - q_n).$$

So we have

$$\sum_{n=1}^{N} p_n^1 - \sum_{n=1}^{N} p_n^1 \frac{q_n}{q_n^1} \geq \sum_{n=1}^{N} (q_n^1 - q_n),$$

or

$$\sum_{n=1}^{N} p_n^1 \frac{q_n}{q_n^1} \leq \sum_{n=1}^{N} p_n^1 - q_n^1 + q_n.$$

Now we write

$$KL(p, p^1) + KL(p, q) - KL(p, q^1) = \sum_{n=1}^{N} \Big(p_n \log \frac{p_n q_n^1}{p_n^1 q_n} + p_n^1 - p_n + q_n - q_n^1 \Big).$$

Using the inequality

$$\log \frac{p_n q_n^1}{p_n^1 q_n} \geq 1 - \frac{p_n^1 q_n}{p_n q_n^1},$$

we have

$$\sum_{n=1}^{N} p_n \Big(\log \frac{p_n q_n^1}{p_n^1 q_n} \Big) \geq \sum_{n=1}^{N} p_n - \sum_{n=1}^{N} p_n^1 \frac{q_n}{q_n^1}$$

$$\geq \sum_{n=1}^{N} p_n - p_n^1 + q_n^1 - q_n.$$

It follows that

$$KL(p, p^1) + KL(p, q) \geq KL(p, q^1),$$

which is the four-point property. It follows from the main theorem that

$$KL(p^n, q^n) \rightarrow \inf_{p \in P,\, q \in Q} KL(p, q) = d.$$

Suppose now that there are vectors $\hat{p} \in P$ and $\hat{q} \in Q$ such that

$$KL(\hat{p}, \hat{q}) \leq KL(p, q),$$

for all $p \in P$ and $q \in Q$. From the three- and four-point properties it follows that

$$KL(\hat{p}, p^n) - KL(\hat{p}, p^{n+1}) \geq KL(p^n, q^n) - KL(\hat{p}, \hat{q}) \geq 0,$$

for all n. Then the sequence $\{KL(\hat{p}, p^n)\}$ is decreasing and

$$\{KL(p^n, q^n)\} \rightarrow KL(\hat{p}, \hat{q}).$$

Since the distance $KL(p, q)$ has bounded level sets in each variable, it follows that the sequence $\{p^n\}$ is bounded. From the four-point property, we have

$$KL(p^*, p^n) + KL(\hat{p}, \hat{q}) \geq KL(p^*, q^n),$$

which tells us that the sequence $\{q^n\}$ is also bounded. Passing to subsequences as needed, we may conclude that there are subsequences $\{p^{n_k}\}$ and $\{q^{n_k}\}$ converging to $p^* \in P$ and $q^* \in Q$, respectively, and that

$$KL(p^*, q^*) = KL(\hat{p}, \hat{q}).$$

Substituting p^* for $\hat{p}$, we find that the sequence $\{KL(p^*, p^n)\}$ is decreasing; since a subsequence converges to zero, the entire sequence converges to zero and $p^n \to p^*$.

5 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 [34]; the book by McLachnan and Krishnan [48] is a good source for the history of this more general algorithm.

It was noticed by Rockmore and Macovski [51] that the image reconstruction problems posed by medical tomography could be formulated as statistical parameter estimation problems. Following up on this idea, Shepp and Vardi [53] suggested the use of the EM algorithm for solving the reconstruction problem in emission tomography. In [46], 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 [53] for the emission case. In [55], Vardi, Shepp and Kaufman repaired the earlier proof, relying on techniques due to Csiszár and Tusnády [32]. In [47] Lange, Bahn and Little improve 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 [9, 10].

The EMML, as a statistical parameter estimation technique, was not originally thought to be connected to any system of linear equations. In [9], it was shown that the EMML algorithm minimizes the function $f(x) = KL(y, Px)$, over nonnegative vectors x. Here y is a vector with positive entries, and P is a matrix with nonnegative entries, such that $s_j = \sum_{i=1}^{I} P_{ij} > 0$. Consequently, when the non-negative system of linear equations $Px = y$ has a non-negative solution, the EMML converges to such a solution.

Because $KL(y, Px)$ is continuous in the variable x and has bounded level sets, there is at least one non-negative minimizer; call it $\hat{x}$. The vector $P\hat{x}$ is unique, even if $\hat{x}$ is not. For convenience, we assume that the problem has been normalized so that $s_j = 1$, for all j.

5.1 The Alt Min Framework for EMML

For each $x \geq 0$, let $q(x)$ and $r(x)$ be the I by J arrays with entries

$$q(x)_{ij} = x_j P_{ij}, \tag{5.1}$$

and

$$r(x)_{ij} = x_j P_{ij} y_i / (Px)_i, \tag{5.2}$$

whenever $q(x)_{ij} > 0$, and $r(x)_{ij} = 0$ otherwise. We then let

$$P = \{r = r(z) | z \geq 0\},$$

and

$$Q = \{q = q(x) | x \geq 0\}.$$

The sets P and Q are closed and convex in the space R^{I+J}. We also define

$$\Theta(p, q) = KL(r(z), q(x)),$$

and

$$\Delta(p, p') = KL(r(z), r(z')),$$

where $x \geq 0$,$z \geq 0$ and $z' \geq 0$ are arbitrary.

5.2 The EMML Iteration

The iterative step of the EMML is to minimize the function $KL(r(x^{n-1}), q(x))$ to get $x = x^n$. The EMML iteration begins with a positive vector x^0. Having found the vector x^{n-1}, the next vector in the EMML sequence is x^n, with entries given by

$$x_j^n = x_j^{n-1} \sum_{i=1}^{I} P_{ij} \Big(\frac{y_i}{(Px^{n-1})_i} \Big). \tag{5.3}$$

It follows from the discussion in the previous section that the sequence $\{x^n\}$ converges to a non-negative minimizer of the function $KL(y, Px)$.

5.3 Pythagorean Identities

We can be a bit more precise about the iterations in the EMML algorithm. We have the following Pythagorean identities [11]:

- **1.** $KL(r(z), q(x)) = KL(r(x), q(x)) + KL(r(z), r(x))$; and
- **2.** $KL(r(z), q(x)) = KL(r(z), q(z')) + KL(z', x)$,

where

$$z'_j = z_j \sum_{i=1}^{I} P_{ij} \frac{y_i}{(Pz)_i},$$

for each j. Note that $KL(y, Px) = KL(r(x), q(x))$.

6 The SMART

What is usually called the *simultaneous multiplicative algebraic reconstruction technique* (SMART) was discovered in 1972, independently, by Darroch and Ratcliff [33], working in statistics, and by Schmidlin [52] in medical imaging. The SMART provides another example of alternating minimization having the three- and four-point properties.

Darroch and Ratcliff called their algorithm *generalized iterative scaling*. It was designed to calculate the entropic projection of one probability vector onto a family of probability vectors with a pre-determined marginal distribution. They did not consider the more general problems of finding a non-negative solution of a non-negative system of linear equations $y = Px$, or of minimizing a function; they did not, therefore, consider what happens in the inconsistent case, in which the system of equations $y = Px$ has no non-negative solutions. This issue was resolved in [9], where it was shown that the SMART minimizes the function $f(x) = KL(Px, y)$, over nonnegative vectors x. This function is continuous in the variable x and has bounded level sets, so there is at least one minimizer; call it $\hat{x}$. The vector $P\hat{x}$ is unique, even if the vector $\hat{x}$ is not. Again, y is a vector with positive entries, and P is a matrix with nonnegative entries, such that $s_j = \sum_{i=1}^{I} P_{ij} = 1$.

6.1 SMART as Alt Min

To put the SMART algorithm into the framework of alternating minimization, we take the sets P and Q as in the EMML case, but now let $p^n = q(x^n)$, and $q^n = r(x^n)$. Generic vectors are $p = q(z)$ for some z and $q = r(x)$ for some x. Then we set

$$\Theta(p, q) = KL(q(x), r(z)),$$

and, for arbitrary $p = q(z)$ and $p' = q(w)$,

$$\Delta(p, p') = KL(q(z), q(w)) = KL(z, w).$$

Note that $KL(Px, y) = KL(q(x), r(x))$.

6.2 The SMART Iteration

The iterative step of the SMART is to minimize the function $KL(q(x), r(x^{n-1}))$ to get $x = x^n$. The SMART iteration begins with a positive vector x^0. Having found the vector x^{n-1}, the next vector in the SMART sequence is x^n, with entries given by

$$x_j^n = x_j^{n-1} \exp\Big(\sum_{i=1}^{I} P_{ij} \log\Big(\frac{y_i}{(Px^{n-1})_i}\Big)\Big). \tag{6.1}$$

It follows from our discussion of entropic projection onto convex sets that the sequence $\{x^n\}$ converges to a non-negative minimizer of the function $KL(Px, y)$.

The Pythagorean identities in the SMART case are more helpful than in the EMML case and enable us to prove more about the SMART.

6.3 The Pythagorean Identities

With $r(z)$ and $q(x)$ defined as above, we have the following Pythagorean identities:

- **1.** $KL(q(x), r(z)) = KL(q(x), r(x)) + KL(x, z) - KL(Px, Pz)$; and
- **2.** $KL(q(x), r(z)) = KL(q(z'), r(z)) + KL(x, z')$,

where

$$z'_j = z_j \exp\Big(\sum_{i=1}^{I} P_{ij} \log\Big(\frac{y_i}{(Pz)_i}\Big)\Big),$$

for each j. From the Pythagorean identity

$$KL(q(x), r(z)) = KL(q(z'), r(z)) + KL(x, z')$$

we have

$$\Theta(p, q^n) = \Theta(p^{n+1}, q^n) + \Delta(p, p^{n+1}),$$

which is then the three-point property for p.

6.4 Convergence of the SMART

Using the three- and four-point properties, we are able to show that

$$KL(\hat{x}, x^{n-1}) - KL(\hat{x}, x^n) \geq KL(q(x^n), r(x^n)) - KL(q(\hat{x}), r(\hat{x})) \geq 0,$$

so that we now know that

$$KL(Px^n, y) = KL(q(x^n), r(x^n)) \to KL(q(\hat{x}), r(\hat{x})).$$

We also know from the fact that

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

for $n = 1, 2, ...$, that the sequence $\{x^n\}$ is bounded. Therefore, there is a cluster point x^* that is the limit of a subsequence $\{x^{n_k}\}$ of $\{x^n\}$. From

$$\{KL(q(x^{n_k}), r(x^{n_k}))\} \to KL(q(x^*), r(x^*)),$$

we conclude that

$$KL(q(x^*), r(x^*)) = KL(q(\hat{x}), r(\hat{x})),$$

so that x^* is a minimizer of $KL(Px, y)$. Therefore, replacing $\hat{x}$ with x^*, we learn that the sequence $\{KL(x^*, x^n)\}$ is decreasing; but a subsequence converges to zero, so the entire sequence converges to zero. Therefore, $\{x^n\}$ converges to x^*. We can actually show even more.

Instead of just

$$KL(\hat{x}, x^{n-1}) - KL(\hat{x}, x^n) \geq KL(q(x^n), r(x^n)) - KL(q(\hat{x}), r(\hat{x})) \geq 0,$$

we can use the Pythagorean identities to obtain

$$KL(\hat{x}, x^{n-1}) - KL(\hat{x}, x^n) = KL(q(x^n), r(x^n)) - KL(q(\hat{x}), r(\hat{x}))$$

$$+KL(P\hat{x}, Px^{n-1}) + KL(x^n, x^{n-1}) - KL(Px^n, Px^{n-1}).$$

This tells us that $KL(\hat{x}, x^{n-1}) - KL(\hat{x}, x^n)$ depends on the vector $P\hat{x}$, but is otherwise independent of the choice of $\hat{x}$. Consequently, by summing both sides of the equation over the index n, we find that $KL(\hat{x}, x^0) - KL(\hat{x}, x^*)$ is also independent of the choice of $\hat{x}$. Therefore, minimizing $KL(x, x^0)$ over all $x \geq 0$ that minimize $KL(Px, y)$ is equivalent to minimizing $KL(x, x^*)$ over all such x; but the answer to the latter problem is clearly $x = x^*$. Therefore, $x = x^*$ is the minimizer of $KL(x, x^0)$ over all $x \geq 0$ that minimize $KL(Px, y)$.

6.5 Related work of Csiszár

In [31] Csiszár shows that the generalized iterative scaling method of Darroch and Ratcliff can be formulated in terms of successive entropic projection onto the sets P

and Q. In other words, he views their method as an alternating projection method, not as alternating minimization. He derives the generalized iterative scaling algorithm in two steps:

- 1. minimize $KL(r(x), q(x^n))$ to get $r(x^n)$; and then
- 2. minimize $KL(q(x), r(x^n))$ to get $q(x^{n+1})$.

Although [31] appeared five years after [32], Csiszár does not reference [32], nor does he mention alternating minimization, instead basing his convergence proof here on his earlier paper [30], which deals with entropic projection. He is able to make this work because the order of the $q(x^n)$ and $r(x)$ does not matter in the first step. Therefore, the generalized iterative scaling, and, more generally, the SMART, is also an alternating minimization algorithm, as well.

7 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 the examples presented above, the function $\Theta(p, q)$ is a distance between p and q, either $\|p-q\|^2$ or $KL(p, q)$. In the case of $\Theta(p, q) = \|p-q\|^2$, each step of the alternating minimization algorithm involves an orthogonal projection onto a closed convex set; both projections are with respect to the same Euclidean distance function. In the case of cross-entropy minimization, we first project q^n onto the set P by minimizing the distance $KL(p, q^n)$ over all $p \in P$, and then project p^{n+1} onto the set Q by minimizing the distance function $KL(p^{n+1}, q)$. This suggests the possibility of using alternating minimization with respect to more general distance functions. We shall focus on Bregman distances.

7.1 Bregman Distances

Let $f : R^N \to R$ be a Bregman function [6, 24, 8], and so $f(x)$ is convex on its domain and differentiable in the interior of its domain. Then, for x in the domain and z in the interior, we define the Bregman distance $D_f(x, z)$ by

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

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

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

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

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

for all $p \in P$. Since

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

it follows that the three-point property holds, with

$$\Theta(p, q) = D_f(p, q),$$

and

$$\Delta(p, p') = D_f(p, p').$$

To get the four-point property we need to restrict D_f somewhat; we assume from now on that $D_f(p, q)$ is jointly convex, that is, it is convex in the combined vector variable (p, q) (see [3]). Now we can invoke a lemma due to Eggermont and LaRiccia [37].

7.2 The Eggermont-LaRiccia Lemma

Lemma 7.1 *Suppose that the Bregman distance $D_f(p, q)$ is jointly convex. Then it has the four-point property.*

Proof: By joint convexity we have

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

$$\langle \nabla_1 D_f(p^n, q^n), p - p^n \rangle + \langle \nabla_2 D_f(p^n, q^n), q - q^n \rangle,$$

where ∇_1 denotes the gradient with respect to the first vector variable. Since q^n minimizes $D_f(p^n, q)$ over all $q \in Q$, we have

$$\langle \nabla_2 D_f(p^n, q^n), q - q^n \rangle \geq 0,$$

for all q. Also,

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

It follows that

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

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

Therefore, we have

$$D_f(p,p^n) + D_f(p,q) \geq D_f(p,q^n).$$

This is the four-point property. ∎

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

7.3 Minimizing a Proximity Function

We present now an example of alternating Bregman distance minimization, taken from [14]. The problem is the *convex feasibility problem* (CFP), to find a member of the intersection $C \subseteq R^J$ of finitely many closed convex sets C_i, $i = 1, ..., I$, or, failing that, to minimize the proximity function

$$F(x) = \sum_{i=1}^{I} D_i(\overleftarrow{P}_i x, x), \tag{7.1}$$

where f_i are Bregman functions for which D_i, the associated Bregman distance, is jointly convex, and $\overleftarrow{P}_i x$ the *backward* Bregman projection of x onto the set C_i, that is,

$$D_i(\overleftarrow{P}_i x, x) \leq D_i(z, x),$$

for all $z \in C_i$. Because each D_i is jointly convex, the function $F(x)$ is convex.

The problem can be formulated as an alternating minimization, where $P \subseteq R^{IJ}$ is the product set

$$P = C_1 \times C_2 \times ... \times C_I$$

so that a typical member of P has the form

$$p = (c_1, c_2, ..., c_I),$$

where $c_i \in C_i$, and $Q \subseteq R^{IJ}$ is the *diagonal* subset, meaning that the elements of Q are the I-fold product of a single x; that is

$$Q = \{d(x) = (x, x, ..., x) \in R^{IJ}\}.$$

We then take

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

and $\Delta(p,p') = \Theta(p,p')$.

In [21] 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.

7.4 Forward and Backward Projections

Because Bregman distances D_f are not generally symmetric, we can speak of *forward* and *backward* Bregman projections onto a closed convex set. For any allowable vector x, the *backward* Bregman projection of x onto C, if it exists, is the vector $\overleftarrow{P}_C x$ satisfying the inequality

$$D_f(\overleftarrow{P}_C x, x) \leq D_f(c, x),$$

for all $c \in C$. Similarly, the *forward* Bregman projection is the vector $\overrightarrow{P}_C x$ satisfying the inequality

$$D_f(x, \overrightarrow{P}_C x) \leq D_f(x, c),$$

for any $c \in C$.

The alternating minimization approach described above to minimize the proximity function

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

can be viewed as an alternating projection method, but employing both forward and backward Bregman projections.

Consider the problem of finding a member of the intersection of two closed convex sets C and D. We could proceed as follows: having found x^n, minimize $D_f(x^n, d)$ over all $d \in D$, obtaining $d = \overrightarrow{P}_D x^n$, and then minimize $D_f(c, \overrightarrow{P}_D x^n)$ over all $c \in C$, obtaining

$$c = x^{n+1} = \overleftarrow{P}_C \overrightarrow{P}_D x^n.$$

The objective of this algorithm is to minimize $D_f(c, d)$ over all $c \in C$ and $d \in D$; such minimizers may not exist, of course.

In [4] the authors note that the alternating minimization algorithm of [14] involves forward and backward Bregman projections, which suggests to them iterative methods involving a wider class of operators that they call "Bregman retractions".

8 More Proximity Function Minimization

Proximity function minimization and forward and backward Bregman projections play a role in a variety of iterative algorithms. We survey several of them in this section.

8.1 Cimmino's Algorithm

Our objective here is to find an exact or approximate solution of the system of I linear equations in J unknowns, written $Ax = b$. For each i let

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

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

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

where

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

Let

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

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

$$x_j^{n+1} = x_j^n + \frac{1}{I} \sum_{i=1}^{I} \alpha_i A_{ij}(b_i - (Ax^n)_i). \tag{8.1}$$

8.2 Simultaneous Projection for Convex Feasibility

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

$$x^{n+1} = \frac{1}{I} \sum_{i=1}^{I} P_i x^n.$$

The objective here is to minimize $F(x)$, if it has a minimizer.

8.3 The EMML Revisited

As in our earlier discussion of the EMML method, we want an exact or approximate solution of the system $y = Px$. For each i, let

$$C_i = \{z \geq 0 \,|(Pz)_i = y_i\}.$$

The backward entropic projection of $x > 0$ onto C_i is the vector that minimizes $KL(c_i, x)$, over all $c_i \in C_i$; unfortunately, we typically cannot calculate this projection in closed form. Instead, we define the distances

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

and calculate the associated backward projections $\overleftarrow{P}_i x$ onto the sets C_i; we have

$$D_i(\overleftarrow{P}_i x, x) \leq D_i(c_i, x),$$

for all $c_i \in C_i$. We can calculate $\overleftarrow{P}_i x$ in closed form:

$$(\overleftarrow{P}_i x)_j = x_j \frac{y_i}{(Px)_i},$$

for each j. Note that, for the distances D_i and these sets C_i, the backward and forward projections are the same; that is

$$\overleftarrow{P}_i x = \overrightarrow{P}_i x.$$

Applying alternating minimization to the proximity function

$$F(x) = \sum_{i=1}^{I} \sum_{j=1}^{J} P_{ij} KL(\overleftarrow{P}_i x, x),$$

we obtain the iterative step

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

which is the EMML iteration.

8.4 The SMART

Now we define the proximity function $F(x)$ to be

$$F(x) = \sum_{i=1}^{I} \sum_{j=1}^{J} P_{ij} KL(x, \overrightarrow{P}_i x).$$

Applying alternating minimization and using the fact that $\overleftarrow{P}_i x = \overrightarrow{P}_i x$, we discover that the resulting iterative step is that of the SMART.

9 Alternating Bregman Projection

The generalized alternating projection method is to minimize $D_1(p, q^n)$ over $p \in P$ to get p^{n+1} and then to minimize $D_2(q, p^{n+1})$ over $q \in Q$ to get q^{n+1}. A more tractable problem is alternating Bregman projection involving two distinct Bregman distances: $D_1 = D_f$ and $D_2 = D_g$, where f and g are Bregman functions.

Now we consider minimizing $D_f(p, q^n)$ over $p \in P$ and then minimizing $D_g(q, p^n)$ over $q \in Q$, where, for simplicity, we assume that both P and Q are subsets of the interiors of both domains. Generally, however, this approach of alternating Bregman projection does not work, as the following example illustrates.

9.1 A Counter-example

Let Q be a closed convex subset of R^N and A an invertible N by N real matrix. We consider the problem of projecting a vector x onto the set $A^{-1}(Q)$. Because finding the orthogonal projection of x onto $A^{-1}(Q)$ is difficult, we choose instead to minimize the function

$$F(x) = (x - A^{-1}q)^T A^T A(x - A^{-1}q),$$

over all $q \in Q$. Since

$$F(x) = \frac{1}{2}\|x - A^{-1}q\|^2_{A^TA} = \frac{1}{2}\|A(x - A^{-1}q)\|^2 = \frac{1}{2}\|Ax - q\|^2,$$

this oblique projection can be written in closed form; it is $A^{-1}P_QAx$. This suggests the possibility of a sequential projection algorithm whereby we first perform the oblique projection of x onto $A^{-1}(Q)$, and then the orthogonal projection of $A^{-1}P_QAx$ onto C. Unfortunately, this approach does not work in general, as the following counter-example illustrates.

Let R^N be R^2, with C the x-axis and Q the y-axis. Let A be the matrix

$$A = \begin{bmatrix} 1 & -1 \\ -1 & 0 \end{bmatrix}.$$

Starting with $x^0 = (1, 0)^T$, we find that $P_CA^{-1}P_QAx^0 = x^0$, so the proposed sequential method does not converge, even though the SFP has a solution, namely $(0, 0)^T$.

10 Multiple-Distances Generalized Projection

We want to use the Bregman distances D_{f_i} and the associated backward Bregman projections onto the C_i, which we denote by $\overleftarrow{P}_i$, to obtain a sequential algorithm for

finding a member of the intersection of the sets C_i. As the counter-example showed, the obvious way of simply taking one projection after another, does not always work. As we shall see, what we need to do is to introduce a certain kind of *relaxation* into every step [13].

Instead of simply taking $\overleftarrow{P}_i x^k = x^{k+1}$, we want x^{k+1} to be some sort of combination of $\overleftarrow{P}_i x^k$ and x^k itself. To achieve this, we want to minimize a sum of two distances,

$$D_{f_i}(x, \overleftarrow{P}_i x^k) + D_i(x, x^k),$$

for some carefully chosen distance D_i. What works is to select a Bregman function h with the property that $D_h(x, z) \geq D_{f_i}(x, z)$ for all i, and let $D_i = D_h - D_{f_i}$. For example, we could take $h(x) = \sum_{i=1}^{I} f_i(x)$.

To get x^{k+1} we minimize

$$D_{f_i}(x, \overleftarrow{P}_i x^k) + D_h(x, x^k) - D_{f_i}(x, x^k).$$

It follows that

$$0 = \nabla f_i(x^{k+1}) - \nabla f_i(\overleftarrow{P}_i x^k) + \nabla h(x^{k+1}) - \nabla h(x^k) - \nabla f_i(x^{k+1}) - \nabla f_i(x^k),$$

so that

$$\nabla h(x^{k+1}) = \nabla f_i(\overleftarrow{P}_i x^k) + \nabla h(x^k) - \nabla f_i(x^k).$$

11 The Split Feasibility Problem

Let C and D be non-empty closed convex subsets of R^N and R^M, respectively, and A a real M by N matrix. The *split feasibility problem* (SFP) [21] is to find $x \in C$ with $Ax \in D$. We can reformulate this problem as alternating Euclidean distance minimization in two ways. We can let $P = C$ and $Q = A^{-1}(D)$ and apply alternating minimization to the distance $\|p - q\|^2$, or we can let $P = A(C)$ and $Q = D$ and minimize $\|p - q\|^2$. While we may wish to assume that the orthogonal projections P_C and P_Q are easy to calculate, as they are for C and Q that are nice enough, it is unlikely that the orthogonal projections onto $A^{-1}(D)$ or $A(C)$ will be easy to compute. This suggests the possibility of using distinct distances for the projections.

11.1 Alternating Oblique Projections

For example, we can let $P = C$ and $Q = A^{-1}(D)$, but when we perform the projection of p onto the set Q, we minimize the weighted Euclidean distance

$$\text{dist}^2(p, q) = (p - q)^T A^T A(p - q);$$

for convenience, we assume that $M = N$ and that A is invertible. Clearly, minimizing $\text{dist}^2(p^n, q)$ with respect to q is equivalent to minimizing $\|Ap^n - d\|^2$, with respect to $d \in D$. The optimal d is $d = P_D(Ap^n)$, so the optimal q is $q = A^{-1}P_DAp^n$. The next p^{n+1} will be $p^{n+1} = P_CA^{-1}P_DAp^n$. But, as we saw in the counter-example above, this iterative procedure need not converge to a solution of the SFP. We want an iterative algorithm that solves the SFP and employs P_C and P_D.

11.2 Projected Gradient Minimization

Suppose that $f : R^N \to R$ is differentiable and we want to minimize $f(x)$ over $x \in R^N$. The gradient descent method has the iterative step

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

where the parameters γ_k are selected to guarantee convergence to a minimizer. If the original problem is actually to minimize $f(x)$, not over all $x \in R^N$, but only over $x \in C$, where C is some non-empty closed convex set, then the *projected gradient descent* method can be employed. The iterative step now is

$$x^{k+1} = P_C(x^k - \gamma_k \nabla f(x^k)).$$

We apply this idea to the SFP.

We try to solve the SFP by minimizing the function

$$f(x) = \|P_DAx - Ax\|^2,$$

subject to $x \in C$. The gradient of $f(x)$ is

$$\nabla f(x) = A^TAx - A^TP_DAx,$$

so the iterative step becomes

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

This algorithm was presented in [15] and discussed further in [16]. In those papers it is called the CQ algorithm. It turns out that, for the CQ algorithm, we get convergence to a solution of the SFP, whenever such exist, for any γ in the interval $(0, 2/L)$, where L is the largest eigenvalue of A^TA. When no solutions exist, the iteration provides a minimizer of the function $f(x)$, for $x \in C$, provided that such minimizers exist.

In practice, the orthogonal projections P_C and P_D often need to be approximated by orthogonal projection onto supporting hyperplanes. The paper by Yang [56] is one of several articles that explore this issue.

12 Multiple-Set Split Feasibility

In intensity modulated radiation therapy (IMRT) the objective is to modulate the intensity levels of the radiation sources around the body of the patient so as to direct sufficient radiation to the intended target while avoiding other regions [19, 20]. There are mechanical constraints on the manner in which the modulation can be performed. Taken together, the problem becomes a split-feasibility problem in which both C and Q are the intersection of finitely many other closed convex subsets.

Let $C_n, n = 1, ..., N$ and $Q_m, m = 1, ..., M$ be closed convex subsets of R^J and R^I, respectively, with $C = \cap_{n=1}^N C_n$ and $Q = \cap_{m=1}^M Q_m$. Let A be a real I by J matrix. The *multiple-sets split feasibility problem* (MSSFP) is to find $x \in C$ with $Ax \in Q$. The assumption is that the orthogonal projections onto each C_n and each Q_m can be easily computed.

A somewhat more general problem is to find a minimizer of the proximity function

$$f(x) = \frac{1}{2}\sum_{n=1}^{N} \alpha_n ||P_{C_n}x - x||_2^2 + \frac{1}{2}\sum_{m=1}^{M} \beta_m ||P_{Q_m}Ax - Ax||_2^2, \tag{12.1}$$

with respect to the nonempty, closed convex set $\Omega \subseteq R^N$, where α_n and β_m are positive and

$$\sum_{n=1}^{N} \alpha_n + \sum_{m=1}^{M} \beta_m = 1.$$

In [20] it is shown that $\nabla f(x)$ is L-Lipschitz, for

$$L = \sum_{n=1}^{N} \alpha_n + \rho(A^T A) \sum_{m=1}^{M} \beta_m;$$

here $\rho(A^T A)$ is the spectral radius of $A^T A$, which is also its largest eigenvalue. The algorithm given in [20] has the iterative step

$$x^{k+1} = P_\Omega\Big(x^k + s\Big(\sum_{n=1}^{N} \alpha_n (P_{C_n}x^k - x^k) + \sum_{m=1}^{M} \beta_m A^T (P_{Q_m}Ax^k - Ax^k)\Big)\Big), \tag{12.2}$$

for $0 < s < 2/L$. This algorithm converges to a minimizer of $f(x)$ over Ω, whenever such a minimizer exists, and to a solution, within Ω, of the MSSFP, whenever such solutions exist.

13 Forward-Backward Splitting

Let $f : R^J \to (-\infty, +\infty]$ be a closed, proper, convex function. When f is differentiable, we can find minimizers of f using techniques such as gradient descent. When

f is not necessarily differentiable, the minimization problem is more difficult. One approach is to augment the function f and to convert the problem into one of minimizing a differentiable function. Moreau's approach uses Euclidean distances to augment f, leading to the definition of *proximal operators* [50], or *proximity operators* [29]. More general methods, using Bregman distances to augment f, have been considered by Teboulle [54] and by Censor and Zenios [23].

13.1 Moreau's Proximity Operators

The Moreau envelope of the function f is the function

$$m_f(z) = \inf_x \{f(x) + \frac{1}{2}||x - z||_2^2\}, \tag{13.1}$$

which is also the infimal convolution of the functions $f(x)$ and $\frac{1}{2}||x||_2^2$. It can be shown that the infimum is uniquely attained at the point denoted $x = \text{prox}_f z$ (see [50]). The function $m_f(z)$ is differentiable and $\nabla m_f(z) = z - \text{prox}_f z$. The point $x = \text{prox}_f z$ is characterized by the property $z - x \in \partial f(x)$. Consequently, x is a global minimizer of f if and only if $x = \text{prox}_f x$. For example, consider the indicator function of the convex set C, $f(x) = \iota_C(x)$ that is zero if x is in the closed convex set C and $+\infty$ otherwise. Then $m_f z$ is the minimum of $\frac{1}{2}||x - z||_2^2$ over all x in C, and $\text{prox}_f z = P_C z$, the orthogonal projection of z onto the set C.

13.2 Proximal Minimization Algorithm

The *proximal minimization algorithm* is the following.

Algorithm 13.1 (Proximal Minimization) *Let x^0 be arbitrary. For $k = 0, 1, ...$, let*

$$x^{k+1} = (1 - \gamma_k)x^k + \gamma_k \text{prox}_f x^k. \tag{13.2}$$

Because

$$x^k - \text{prox}_f x^k \in \partial f(\text{prox}_f x^k), \tag{13.3}$$

the iteration in Equation (13.2) has the increment

$$x^{k+1} - x^k \in -\gamma_k \partial f(x^{k+1}), \tag{13.4}$$

in contrast to what we would have with the usual gradient descent method for differentiable f:

$$x^{k+1} - x^k = -\gamma_k \nabla f(x^k). \tag{13.5}$$

It follows from the definition of the sub-differential $\partial f(x^{k+1})$ that $f(x^k) \geq f(x^{k+1})$ for the iteration in Equation (13.2).

13.3 Forward-Backward Splitting

In [29] Combettes and Wajs consider the problem of minimizing the function $f = f_1 + f_2$, where f_2 is differentiable and its gradient is λ-Lipschitz continuous. The function f is minimized at the point x if and only if

$$0 \in \partial f(x) = \partial f_1(x) + \nabla f_2(x), \tag{13.6}$$

so we have

$$-\gamma \nabla f_2(x) \in \gamma \partial f_1(x), \tag{13.7}$$

for any $\gamma > 0$. Therefore

$$x - \gamma \nabla f_2(x) - x \in \gamma \partial f_1(x). \tag{13.8}$$

From Equation (13.8) we conclude that

$$x = \text{prox}_{\gamma f_1}(x - \gamma \nabla f_2(x)). \tag{13.9}$$

This suggests an algorithm, called the *forward-backward splitting* for minimizing the function $f(x)$.

Algorithm 13.2 (Forward-Backward Splitting) *Beginning with an arbitrary x^0, and having calculated x^k, we let*

$$x^{k+1} = \text{prox}_{\gamma f_1}(x^k - \gamma \nabla f_2(x^k)), \tag{13.10}$$

with γ chosen to lie in the interval $(0, 2/\lambda)$.

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

13.4 The CQ Algorithm as Forward-Backward Splitting

Recall that the split-feasibility problem (SFP) is to find x in C with Ax in QD. The CQ algorithm minimizes the function

$$g(x) = ||P_D Ax - Ax||_2^2, \tag{13.11}$$

over $x \in C$, whenever such minimizers exist, and so solves the SFP whenever it has solutions. The CQ algorithm minimizes the function

$$f(x) = \iota_C(x) + g(x), \tag{13.12}$$

where ι_C is the indicator function of the set C. With $f_1(x) = \iota_C(x)$ and $f_2(x) = g(x)$, the function $f(x)$ has the form considered by Combettes and Wajs, and the CQ algorithm becomes a special case of their forward-backward splitting method.

14 Projecting onto the Intersection of Closed Convex Sets

The SOP algorithm need not converge to the point in the intersection closest to the starting point. To obtain the point closest to x^0 in the intersection of the convex sets C_i, we can use *Dykstra's algorithm*, a modification of the SOP method [36]. For simplicity, we shall discuss only the case of $C = A \cap B$, the intersection of two closed, convex sets.

14.1 A Motivating Lemma

The following lemma will help to motivate Dykstra's algorithm.

Lemma 14.1 *If $x = c + p + q$, where $c = P_A(c + p)$ and $c = P_B(c + q)$, then $c = P_C x$.*

Proof: Let d be arbitrary in C. Then

$$\langle c - (c + p), d - c \rangle \geq 0, \tag{14.1}$$

since d is in A, and

$$\langle c - (c + q), d - c \rangle \geq 0, \tag{14.2}$$

since d is in B. Adding the two inequalities, we get

$$\langle -p - q, d - c \rangle \geq 0. \tag{14.3}$$

But

$$-p - q = c - x, \tag{14.4}$$

so

$$\langle c - x, d - c \rangle \geq 0, \tag{14.5}$$

for all d in C. Therefore, $c = P_C x$. ∎

14.2 Dykstra's Algorithm

Dykstra's algorithm is the following:

Algorithm 14.1 (Dykstra) *Let $b_0 = x$, and $p_0 = q_0 = 0$. Then let*

$$a_n = P_A(b_{n-1} + p_{n-1}), \tag{14.6}$$

$$b_n = P_B(a_n + q_{n-1}), \tag{14.7}$$

and define p_n and q_n by

$$x = a_n + p_n + q_{n-1} = b_n + p_n + q_n. \tag{14.8}$$

Uing the algorithm, we construct two sequences, $\{a_n\}$ and $\{b_n\}$, both converging to $c = P_C x$, along with two other sequences, $\{p_n\}$ and $\{q_n\}$. Usually, but not always, $\{p_n\}$ converges to p and $\{q_n\}$ converges to q, so that

$$x = c + p + q, \tag{14.9}$$

with

$$c = P_A(c + p) = P_B(c + q). \tag{14.10}$$

Generally, however, $\{p_n + q_n\}$ converges to $x - c$.

In [6], Bregman considers the problem of minimizing a convex function $f : R^J \to R$ over the intersection of half-spaces, that is, over the set of points x for which $Ax =\geq b$. His approach is a *primal-dual* algorithm involving the notion of projecting onto a convex set, with respect to a generalized distance constructed from f. Such generalized projections have come to be called *Bregman projections*. In [25], Censor and Reich extend Dykstra's algorithm to Bregman projections, and, in [7], Bregman, Censor and Reich show that the extended Dykstra algorithm of [25] is the natural extension of Bregman's primal-dual algorithm to the case of intersecting convex sets.

14.3 The Halpern-Lions-Wittmann-Bauschke Algorithm

There is yet another approach to finding the orthogonal projection of the vector x onto the nonempty intersection C of finitely many closed, convex sets C_i, $i = 1, ..., I$.

Algorithm 14.2 (HLWB) *Let x^0 be arbitrary. Then let*

$$x^{k+1} = t_k x + (1 - t_k) P_i x^k, \tag{14.11}$$

where P_i denotes the orthogonal projection onto C_i, t_k is in the interval $(0, 1)$, and $i = k(\mathrm{mod}\, I) + 1$.

Several authors have proved convergence of the sequence $\{x^k\}$ to $P_C x$, with various conditions imposed on the parameters $\{t_k\}$. As a result, the algorithm is known as the Halpern-Lions-Wittmann-Bauschke (HLWB) algorithm, after the names of several who have contributed to the evolution of the theorem; see also Corollary 2 in Reich's paper [49]. The conditions imposed by Bauschke [1] are $\{t_k\} \to 0$, $\sum t_k = \infty$, and $\sum |t_k - t_{k+I}| < +\infty$. The HLWB algorithm has been extended by Deutsch and Yamada [35] to minimize certain (possibly non-quadratic) functions over the intersection of fixed point sets of operators more general than P_i.

14.4 Dykstra's Algorithm for Bregman Projections

We are concerned now with finding the backward Bregman projection of x onto the intersection C of finitely many closed convex sets, C_i. The problem can be solved by extending Dykstra's algorithm to include Bregman projections.

14.5 A Helpful Lemma

The following lemma helps to motivate the extension of Dykstra's algorithm.

Lemma 14.2 *Suppose that*

$$\nabla f(c) - \nabla f(x) = \nabla f(c) - \nabla f(c+p) + \nabla f(c) - \nabla f(c+q), \tag{14.12}$$

with $c = \overleftarrow{P}^f_A(c+p)$ *and* $c = \overleftarrow{P}^f_B(c+q)$. *Then* $c = \overleftarrow{P}^f_C x$.

Proof: Let d be arbitrary in C. We have

$$\langle \nabla f(c) - \nabla f(c+p), d - c \rangle \geq 0, \tag{14.13}$$

and

$$\langle \nabla f(c) - \nabla f(c+q), d - c \rangle \geq 0. \tag{14.14}$$

Adding, we obtain

$$\langle \nabla f(c) - \nabla f(x), d - c \rangle \geq 0. \tag{14.15}$$

∎

This suggests the following algorithm for finding $c = \overleftarrow{P}^f_C x$, which turns out to be the extension of Dykstra's algorithm to Bregman projections.

Algorithm 14.3 (Bregman-Dykstra) *Begin with $b^0 = x$, $p_0 = q_0 = 0$. Define*

$$b_{n-1} + p_{n-1} = \nabla f^{-1}(\nabla f(b_{n-1}) + r_{n-1}), \tag{14.16}$$

$$a_n = \overleftarrow{P}_A^f(b_{n-1} + p_{n-1}), \tag{14.17}$$

$$r_n = \nabla f(b_{n-1}) + r_{n-1} - \nabla f(a_n), \tag{14.18}$$

$$\nabla f(a_n + q_{n-1}) = \nabla f(a_n) + s_{n-1}, \tag{14.19}$$

$$b_n = \overleftarrow{P}_B^f(a_n + q_{n-1}), \tag{14.20}$$

and

$$s_n = \nabla f(a_n) + s_{n-1} - \nabla f(b_n). \tag{14.21}$$

In place of

$$\nabla f(c + p) - \nabla f(c) + \nabla f(c + q) - \nabla f(c), \tag{14.22}$$

we have

$$[\nabla f(b_{n-1}) + r_{n-1}] - \nabla f(b_{n-1}) + [\nabla f(a_n) + s_{n-1}] - \nabla f(a_n) = r_{n-1} + s_{n-1}, \tag{14.23}$$

and also

$$[\nabla f(a_n) + s_{n-1}] - \nabla f(a_n) + [\nabla f(b_n) + r_n] - \nabla f(b_n) = r_n + s_{n-1}. \tag{14.24}$$

But we also have

$$r_{n-1} + s_{n-1} = \nabla f(x) - \nabla f(b_{n-1}), \tag{14.25}$$

and

$$r_n + s_{n-1} = \nabla f(x) - \nabla f(a_n). \tag{14.26}$$

Then the sequences $\{a_n\}$ and $\{b_n\}$ converge to c. For further details, see the papers of Censor and Reich [25] and Bauschke and Lewis [5].

In [7] Bregman, Censor and Reich show that the extension of Dykstra's algorithm to Bregman projections can be viewed as an extension of Bregman's primal-dual algorithm to the case in which the intersection of half-spaces is replaced by the intersection of closed convex sets.

15 Acceleration

Some iterative algorithms that find exact or approximate solutions of systems of linear equations, such as the Landweber, the SMART and the EMML methods, tend to be slow to converge. Since the intent is often to use these algorithms to solve large systems of linear equations, speed is important and these methods become impractical.

The Landweber, SMART and EMML algorithms use all of the equations at each step of the iteration and are therefore called *simultaneous* methods. Variants of these methods that use only some of the equations at each step are called *block-iterative* methods. Those employing only a single equation at each step, such as the *algebraic reconstruction technique* (ART) [38, 39], its multiplicative cousin MART [38], and the related EMART, are called *sequential* methods [17]. For problems such as medical image reconstruction, block-iterative methods have been shown to provide useful reconstructed images in a fraction of the time required by the simultaneous methods.

15.1 Block-iterative Versions of SMART and EMML

Darroch and Ratcliff included what are now called block-iterative versions of SMART in their original paper [33]. Censor and Segman [22] viewed SMART and its block-iterative versions as natural extension of the MART. Consequently, block-iterative variants of SMART have been around for some time. The story with the EMML is quite different.

The paper of Holte, Schmidlin, *et al.* [41] compares the performance of Schmidlin's method of [52] with the EMML algorithm. Almost as an aside, they notice the accelerating effect of what they call *projection interleaving*, that is, the use of blocks. This paper contains no explicit formulas, however, and presents no theory, so one can only make educated guesses as to the precise iterative methods employed. Somewhat later, Hudson, Hutton and Larkin [42, 43] observed that the EMML can be significantly accelerated if, at each step, one employs only some of the data. They referred to this approach as the *ordered subset* EM method (OSEM). They gave a proof of convergence of the OSEM, for the consistent case. The proof relied on a fairly restrictive relationship between the matrix A and the choice of blocks, called *subset balance*. In [12] a revised version of the OSEM, called the *rescaled block-iterative* EMML (RBI-EMML), was shown to converge, in the consistent case, regardless of the choice of blocks.

15.2 Basic assumptions

Methods based on cross-entropy, such as the MART, SMART, EMML and all block-iterative versions of these algorithms apply to nonnegative systems that we denote by $Px = y$, where y is a vector of positive entries, P is a matrix with entries $P_{ij} \geq 0$ such that for each j the sum $s_j = \sum_{i=1}^{I} P_{ij}$ is positive and we seek a solution x with nonnegative entries. If no nonnegative x satisfies $y = Px$ we say the system is *inconsistent*.

Simultaneous iterative algorithms employ all of the equations at each step of the iteration; block-iterative methods do not. For the latter methods we assume that the index set $\{i = 1, ..., I\}$ is the (not necessarily disjoint) union of the N sets or *blocks* B_n, $n = 1, ..., N$. We shall require that $s_{nj} = \sum_{i \in B_n} P_{ij} > 0$ for each n and each j. Block-iterative methods like ART and MART for which each block consists of precisely one element are called *row-action* or *sequential* methods. For each $n = 1, ..., N$ let

$$m_n = \max\{s_{nj} s_j^{-1} | j = 1, ..., J\}. \tag{15.1}$$

The original RBI-SMART is as follows:

Algorithm 15.1 (RBI-SMART) *Let x^0 be an arbitrary positive vector. For $k = 0, 1, ...$, let $n = k(\mathrm{mod}\, N) + 1$. Then let*

$$x_j^{k+1} = x_j^k \exp\Big(m_n^{-1} s_j^{-1} \sum_{i \in B_n} P_{ij} \log\Big(\frac{y_i}{(Px^k)_i}\Big)\Big). \tag{15.2}$$

Notice that Equation (15.2) can be written as

$$\log x_j^{k+1} = (1 - m_n^{-1} s_j^{-1} s_{nj}) \log x_j^k + m_n^{-1} s_j^{-1} \sum_{i \in B_n} P_{ij} \log\Big(x_j^k \frac{y_i}{(Px^k)_i}\Big), \tag{15.3}$$

from which we see that x_j^{k+1} is a weighted geometric mean of x_j^k and the terms

$$(\overleftarrow{P}_i x^k)_j = x_j^k \Big(\frac{y_i}{(Px^k)_i}\Big),$$

for $i \in B_n$. This will be helpful in deriving block-iterative versions of the EMML algorithm. The vectors $\overleftarrow{P}_i(x^k)$ are sometimes called weighted KL projections.

For the row-action version of SMART, the *multiplicative* ART (MART), due to Gordon, Bender and Herman [38], we take $N = I$ and $B_n = B_i = \{i\}$ for $i = 1, ..., I$. The MART has the iterative

$$x_j^{k+1} = x_j^k \Big(\frac{y_i}{(Px^k)_i}\Big)^{m_i^{-1} P_{ij}}, \tag{15.4}$$

for $j = 1, 2, ..., J$, $i = k(\mathrm{mod}\, I) + 1$ and $m_i > 0$ chosen so that $m_i^{-1} P_{ij} \leq 1$ for all j. The smaller m_i is the faster the convergence, so a good choice is $m_i = \max\{P_{ij}|,\ j = 1, ..., J\}$. Although this particular choice for m_i is not explicitly mentioned in the various discussions of MART I have seen, it was used in implementations of MART from the beginning [40].

Darroch and Ratcliff included a discussion of a block-iterative version of SMART in their 1972 paper [33]. Close inspection of their version reveals that they require that $s_{nj} = \sum_{i \in B_n} P_{ij} = 1$ for all j. Since this is unlikely to be the case initially, we might try to rescale the equations or unknowns to obtain this condition. However, unless $s_{nj} = \sum_{i \in B_n} P_{ij}$ depends only on j and not on n, which is the *subset balance* property used in [43], we cannot redefine the unknowns in a way that is independent of n.

The MART fails to converge in the inconsistent case. What is always observed, but for which no proof exists, is that, for each fixed $i = 1, 2, ..., I$, as $m \to +\infty$, the MART subsequences $\{x^{mI+i}\}$ converge to separate limit vectors, say $x^{\infty,i}$. This *limit cycle* $\mathrm{LC} = \{x^{\infty,i} | i = 1, ..., I\}$ reduces to a single vector whenever there is a nonnegative solution of $y = Px$. The greater the minimum value of $KL(Px, y)$ the more distinct from one another the vectors of the limit cycle are. An analogous result is observed for BI-SMART.

The original motivation for the RBI-EMML came from consideration of Equation (15.3), replacing the geometric means with arithmetic means. This RBI-EMML is as follows:

Algorithm 15.2 (RBI-EMML) *Let x^0 be an arbitrary positive vector. For $k = 0, 1, ...,$ let $n = k(\mathrm{mod}\, N) + 1$. Then let*

$$x_j^{k+1} = (1 - m_n^{-1} s_j^{-1} s_{nj}) x_j^k + m_n^{-1} s_j^{-1} x_j^k \sum_{i \in B_n} (P_{ij} \frac{y_i}{(Px^k)_i}). \tag{15.5}$$

Both the RBI-SMART and the RBI-EMML converge to a non-negative solution of $y = Px$, when such solutions exist.

References

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

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

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

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

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

[6] Bregman, L.M. (1967) "The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming." *USSR Computational Mathematics and Mathematical Physics* **7**, pp. 200–217.

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

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

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

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

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

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

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

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

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

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

[17] Byrne, C. (2009) "Block-iterative algorithms." *International Transactions in Operations Research* **16(4)**, pp. 427–463.

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

[19] Censor, Y., Bortfeld, T., Martin, B., and Trofimov, A. (2006) "A unified approach for inversion problems in intensity-modulated radiation therapy." *Physics in Medicine and Biology* 51, 2353-2365.

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

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

[22] Censor, Y. and Segman, J. (1987) "On block-iterative maximization." *J. of Information and Optimization Sciences* **8**, pp. 275–291.

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

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

[25] Censor, Y., and Reich, S. (1998) "The Dykstra algorithm for Bregman projections." *Communications in Applied Analysis* **2**, pp. 323–339.

[26] Censor, Y., Gordon, D., and Gordon, R. (2001) "Component averaging: an efficient iterative parallel algorithm for large and sparse unstructured problems." *Parallel Computing* **27**, pp. 777–808.

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

[28] Cimmino, G. (1938) "Calcolo approssimato per soluzioni dei sistemi di equazioni lineari." *La Ricerca Scientifica XVI, Series II, Anno IX* **1**, pp. 326–333.

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

[30] Csiszár, I. (1975) "I-divergence geometry of probability distributions and minimization problems." *The Annals of Probability* **3(1)**, pp. 146–158.

[31] Csiszár, I. (1989) "A geometric interpretation of Darroch and Ratcliff's generalized iterative scaling." *The Annals of Statistics* **17(3)**, pp. 1409–1413.

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

[33] Darroch, J. and Ratcliff, D. (1972) "Generalized iterative scaling for log-linear models." *Annals of Mathematical Statistics* **43**, pp. 1470–1480.

[34] Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977) "Maximum likelihood from incomplete data via the EM algorithm." *Journal of the Royal Statistical Society, Series B* **37**, pp. 1–38.

[35] Deutsch, F., and Yamada, I. (1998) "Minimizing certain convex functions over the intersection of the fixed point sets of non-expansive mappings." *Numerical Functional Analysis and Optimization* **19**, pp. 33–56.

[36] Dykstra, R. (1983) "An algorithm for restricted least squares regression." *J. Amer. Statist. Assoc.* **78 (384)**, pp. 837–842.

[37] Eggermont, P., and LaRiccia, V. (2001) *Maximum Penalized Likelihood Estimation*. New York: Springer.

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

[39] Herman, G. T. and Meyer, L. (1993) "Algebraic reconstruction techniques can be made computationally efficient." *IEEE Transactions on Medical Imaging* **12**, pp. 600–609.

[40] Herman, G. T. (1999) *private communication*.

[41] Holte, S., Schmidlin, P., Linden, A., Rosenqvist, G. and Eriksson, L. (1990) "Iterative image reconstruction for positron emission tomography: a study of convergence and quantitation problems." *IEEE Transactions on Nuclear Science* **37**, pp. 629–635.

[42] Hudson, M., Hutton, B., and Larkin, R. (1992) "Accelerated EM reconstruction using ordered subsets." *Journal of Nuclear Medicine* **33**, p.960.

[43] Hudson, H.M. and Larkin, R.S. (1994) "Accelerated image reconstruction using ordered subsets of projection data." *IEEE Transactions on Medical Imaging* **13**, pp. 601–609.

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

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

[46] Lange, K. and Carson, R. (1984) "EM reconstruction algorithms for emission and transmission tomography." *Journal of Computer Assisted Tomography* **8**, pp. 306–316.

[47] Lange, K., Bahn, M. and Little, R. (1987) "A theoretical study of some maximum likelihood algorithms for emission and transmission tomography." *IEEE Trans. Med. Imag.* **MI-6(2)**, pp. 106–114.

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

[49] Reich, S. (1980) "Strong convergence theorems for resolvents of accretive operators in Banach spaces." *Journal of Mathematical Analysis and Applications*, pp. 287–292.

[50] Rockafellar, R. (1970) *Convex Analysis*. Princeton, NJ: Princeton University Press.

[51] Rockmore, A., and Macovski, A. (1976) "A maximum likelihood approach to emission image reconstruction from projections." *IEEE Transactions on Nuclear Science* **NS-23**, pp. 1428–1432.

[52] Schmidlin, P. (1972) "Iterative separation of sections in tomographic scintigrams." *Nucl. Med.* **15(1)**.

[53] Shepp, L., and Vardi, Y. (1982) "Maximum likelihood reconstruction for emission tomography." *IEEE Transactions on Medical Imaging* **MI-1**, pp. 113–122.

[54] Teboulle, M. (1992) "Entropic proximal mappings with applications to nonlinear programming." *Mathematics of Operations Research* **17(3)**, pp. 670–690.

[55] Vardi, Y., Shepp, L.A. and Kaufman, L. (1985) "A statistical model for positron emission tomography." *Journal of the American Statistical Association* **80**, pp. 8–20.

[56] Yang, Q. (2004) "The relaxed CQ algorithm solving the split feasibility problem." *Inverse Problems* **20**, pp. 1261–1266.