Alternating Minimization, Proximal Minimization and Optimization Transfer Are Equivalent

Charles L. Byrne and Jong Soo Lee
Department of Mathematical Sciences
University of Massachusetts Lowell
Lowell, MA 01854

May 13, 2016

Abstract

Let X be an arbitrary nonempty set and $f : X \to \mathbb{R}$. The objective is to minimize $f(x)$ over $x \in X$. The iterative algorithms considered here are "descent"algorithms, so that $\{f(x^k)\} \downarrow \beta^* \geq -\infty$. We want $\beta^* = \beta \doteq \inf_{x \in X} f(x)$.

In proximal minimization algorithms (PMA) we minimize $f(x) + d(x, x^{k-1})$ to get x^k. The mapping $d : X \times X \to \mathbb{R}_+$ is a "distance"function, with $d(x, x) = 0$ for all x.

In majorization minimization (MM), also called optimization transfer, a second "majorizing" function $g(x|z)$ is introduced, with the properties $g(x|z) \geq f(x)$, for all x and z in X, and $g(x|x) = f(x)$. We then minimize $g(x|x^{k-1})$ to get x^k.

Let $\Phi : X \times Y \to \mathbb{R}_+$, where X and Y are arbitrary nonempty sets. The objective in alternating minimization (AM) is to find $\hat{x} \in X$ and $\hat{y} \in Y$ such that $\Phi(\hat{x}, \hat{y}) \leq \Phi(x, y)$ for all $x \in X$ and $y \in Y$. For each k we minimize $\Phi(x, y^{k-1})$ to get x^{k-1} and then minimize $\Phi(x^{k-1}, y)$ to get y^k. For each $x \in X$, let $y(x) \in Y$ be such that $\Phi(x, y) \geq \Phi(x, y(x))$, for all $y \in Y$; then $y^k = y(x^{k-1})$. Minimizing $\Phi(x, y)$ over all $x \in X$ and $y \in Y$ is equivalent to minimizing $f(x) \doteq \Phi(x, y(x))$ over all $x \in X$. With $d(x, z) \doteq \Phi(x, y(z)) - \Phi(x, y(x))$, minimizing $\Phi(x, y^k)$ is equivalent to minimizing $f(x) + d(x, x^{k-1})$. Therefore, AM, MM, and PMA are equivalent. Each type of algorithm leads to a decreasing sequence $\{f(x^k)\}$.

New conditions on PMA that imply $\beta^* = \beta$ are given, which lead to new conditions on AM for the sequence $\{\Phi(x^k, y^k)\}$ to converge to $\inf_{x,y} \Phi(x, y)$. These conditions can then be translated into the language of MM. Examples are given of each type of algorithm and some open questions are posed.

Key Words: Alternating minimization, optimization transfer, proximal minimization, Bregman distance, convex functions.

2000 Mathematics Subject Classification: Primary 65F10, 65K10; Secondary 90C26, 26B25. **To appear in JNCA**

1 Introduction

Let X be an arbitrary nonempty set and $f : X \to \mathbb{R}$. The objective is to minimize $f(x)$ over $x \in X$. The iterative algorithms considered here are "descent" algorithms, so that $\{f(x^k)\} \downarrow \beta^* \geq -\infty$. We want $\beta^* = \beta \doteq \inf_{x \in X} f(x)$.

In proximal minimization algorithms (PMA) [12, 13] we minimize $f(x)+d(x, x^{k-1})$ to get x^k. The $d : X \times X \to \mathbb{R}_+$ is a "distance" function, with $d(x, x) = 0$, for all x. In majorization minimization (MM), also called optimization transfer, a second "majorizing" function $g(x|z)$ is introduced, with the properties $g(x|z) \geq f(x)$, for all x and z in X, and $g(x|x) = f(x)$. We then minimize $g(x|x^{k-1})$ to get x^k. With

$$d(x, z) \doteq g(x|z) - f(x),$$

it is clear that MM is equivalent to PMA; alternating minimization (AM) algorithms appear to be more general.

Let $\Phi : X \times Y \to \mathbb{R}_+$, where X and Y are arbitrary nonempty sets. The objective in AM is to find $\hat{x} \in X$ and $\hat{y} \in Y$ such that

$$\Phi(\hat{x}, \hat{y}) \leq \Phi(x, y),$$

for all $x \in X$ and $y \in Y$. For each k we minimize $\Phi(x, y^{k-1})$ to get x^{k-1} and then minimize $\Phi(x^{k-1}, y)$ to get y^k. We have the following proposition:

Proposition 1.1 *The AM, PMA, and MM methods are equivalent.*

Proof: We reformulate AM as a method for minimizing a function $f(x)$ of the single variable $x \in X$. For each $x \in X$, let $y(x) \in Y$ be such that $\Phi(x, y) \geq \Phi(x, y(x))$, for all $y \in Y$. Then minimizing $\Phi(x, y)$ over all $x \in X$ and $y \in Y$ is equivalent to minimizing $f(x) \doteq \Phi(x, y(x))$ over all $x \in X$. Every MM algorithm, and therefore every PMA, can be viewed as an application of alternating minimization: define $\Phi(x, z) \doteq g(x|z)$. Minimizing $g(x|x^{k-1})$ to get x^k is equivalent to minimizing $\Phi(x, x^{k-1})$, while minimizing $g(x^k|z)$ is equivalent to minimizing $\Phi(x^k, z)$ and yields $z = x^k$. ∎

Note that $\Phi(x^{k-1}, y^k) = f(x^{k-1})$. The sequence $\{f(x^k)\}$ is decreasing to some β^*.

Each of the algorithms we consider can be reformulated as minimizing some objective function $f(x)$ and can be described by saying that at each step we minimize

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

where $g_k(x) \geq 0$ and $g_k(x^{k-1}) = 0$. Such methods are called *auxiliary-function* (AF) algorithms [7]. For AF algorithms we know that the sequence $\{f(x^k)\}$ is decreasing

to some number $\beta^* \geq -\infty$. If an AF algorithm is in the subclass of SUMMA2 algorithms, then we know that $\beta^* = \beta \doteq \inf_x f(x)$. The Euclidean and Kullback-Leibler distances yield algorithms in the SUMMA2 class, and we suspect that the methods based on the Hellinger and Pearson ϕ^2 distances are also in the SUMMA2 class. Conditions are presented that are sufficient for PMA to be in the SUMMA2 class, and therefore, for $\beta^* = \beta$ for AM, PMA, and MM algorithms. We also consider the use of alternating minimization of distances to obtain approximate solutions of systems of linear equations. The distances considered include the Euclidean, the Kullback-Leibler, the Hellinger, and the Pearson ϕ^2 distances.

2 Auxiliary-Function Methods in Optimization

Let $f : X \to \mathbb{R}$, where X is an arbitrary nonempty set. In applications the set X will have additional structure, but not always that of a Euclidean space; for that reason, it is convenient to impose no structure at the outset. An iterative procedure for minimizing $f(x)$ over $x \in X$ is called an *auxiliary-function* (AF) algorithm [7] if, at each step, we minimize

$$G_k(x) = f(x) + g_k(x), \tag{2.1}$$

where $g_k(x) \geq 0$, and $g_k(x^{k-1}) = 0$. It follows easily that the sequence $\{f(x^k)\}$ is decreasing, so $\{f(x^k)\} \downarrow \beta^* \geq -\infty$. We want more, however; we want $\beta^* = \beta \doteq \inf_{x \in X} f(x)$. To have this we need to impose an additional condition on the auxiliary functions $g_k(x)$; the SUMMA Inequality [7] is one such additional condition.

2.1 The SUMMA Class

We say that an AF algorithm is in the SUMMA class if the SUMMA Inequality holds for all x in X:

$$G_k(x) - G_k(x^k) \geq g_{k+1}(x). \tag{2.2}$$

One consequence of the SUMMA Inequality is

$$g_k(x) + f(x) \geq g_{k+1}(x) + f(x^k), \tag{2.3}$$

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

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

for all k. The sequence $\{g_k(z)\}$ would then be a decreasing sequence of nonnegative terms with the sequence of its successive differences bounded below by $\beta^* - f(z) > 0$.

There are many iterative algorithms that satisfy the SUMMA Inequality [7], such as barrier-function methods [22], and are therefore in the SUMMA class. However, some important methods that are not in this class still have $\beta^* = \beta$; one example is the proximal minimization method of Auslender and Teboulle [2]. This suggests that the SUMMA class, large as it is, is still unnecessarily restrictive. This leads us to the definition of the SUMMA2 class.

2.2 The SUMMA2 Class

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

$$h_k(x) + f(x) \geq h_{k+1}(x) + f(x^k). \tag{2.4}$$

Any algorithm in the SUMMA class is in the SUMMA2 class; use $h_k = g_k$. As in the SUMMA case, we must have $\beta^* = \beta$, since otherwise the successive differences of the sequence $\{h_k(z)\}$ would be bounded below by $\beta^* - f(z) > 0$. It is helpful to note that the functions h_k need not be the g_k, and we do not require that $h_k(x^{k-1}) = 0$. The proximal minimization method of Auslender and Teboulle is in the SUMMA2 class, as is the expectation maximization maximum likelihood (EMML) algorithm [28, 29, 4].

3 PMA is MM

In proximal minimization algorithms (PMA) we minimize

$$f(x) + d(x, x^{k-1}) \tag{3.1}$$

to get x^k. Here $d(x, z) \geq 0$ and $d(x, x) = 0$, so we say that $d(x, z)$ is a distance.

In [14] the authors review the use, in statistics, of "majorization minimization" (MM), also called "optimization transfer". In numerous papers [21, 1] Jeff Fessler and his colleagues use the terminology "surrogate-function minimization" to describe optimization transfer. The objective is to minimize $f : X \to \mathbb{R}$. In MM methods a second "majorizing" function $g(x|z)$ is introduced, with the properties $g(x|z) \geq f(x)$, for all x and z in X, and $g(x|x) = f(x)$. We then minimize $g(x|x^{k-1})$ to get x^k.

Defining

$$d(x,z) \doteq g(x|z) - f(x),$$

it is clear that $d(x,z)$ is a distance and so MM is equivalent to PMA.

4 PMA with Bregman Distances (PMAB)

Let $f : \mathbb{R}^J \to \mathbb{R}$ and $h : \mathbb{R}^J \to \mathbb{R}$ both be convex and differentiable. Let

$$D_h(x,z) \doteq h(x) - h(z) - \langle \nabla h(z), x - z \rangle$$

be the Bregman distance associated with h. At the kth step of a proximal minimization algorithm with Bregman distance (PMAB) we minimize

$$G_k(x) = f(x) + D_h(x, x^{k-1}) = f(x) + h(x) - h(x^{k-1}) - \langle \nabla h(x^{k-1}), x - x^{k-1} \rangle \quad (4.2)$$

to get x^k. It was shown in [7] that

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

so that all PMAB are in the SUMMA class.

In order to minimize $G_k(x)$ we need to solve the equation

$$0 = \nabla f(x) + \nabla h(x) - \nabla h(x^{k-1}) \quad (4.3)$$

for $x = x^k$; generally, this is not easy. Here is a "trick" that can be used to simplify the calculations. Select a function g so that $h \doteq g - f$ is convex and differentiable and so that the equation

$$0 = \nabla g(x) - \nabla g(x^{k-1}) + \nabla f(x^{k-1}) \quad (4.4)$$

is easily solved. As an example, we use this "trick" to derive a gradient descent algorithm and the Landweber algorithm.

5 Gradient Descent and the Landweber Algorithm

Suppose that we want to minimize a convex differentiable function $f : \mathbb{R}^J \to \mathbb{R}$. If the gradient of f, ∇f, is a L-Lipschitz continuous operator, that is, if

$$\|\nabla f(x) - \nabla f(z)\| \leq L\|x - z\|,$$

then the function

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

is convex, for $0 < \gamma \leq 1/L$. For each k we minimize

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

to get x^k. We then have

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

which is a gradient descent algorithm. As a special case we get Landweber's algorithm.

Suppose we want to find a minimizer of the function $f(x) = \|Ax - b\|^2$, where A is a real I by J matrix. Let $g(x) = \frac{1}{\gamma}\|x\|^2$, for some γ in the interval $(0, \frac{1}{L})$, where $L = \rho(A^T A)$, the largest eigenvalue of the matrix $A^T A$. Then the function $h \doteq g - f$ is convex and differentiable. We have

$$D_f(x, y) = \|Ax - Ay\|^2, \tag{5.5}$$

so that

$$D_h(x, y) = \frac{1}{\gamma}\|x - y\|^2 - \|Ax - Ay\|^2. \tag{5.6}$$

At the kth step we differentiate

$$\|Ax - b\|^2 + \frac{1}{\gamma}\|x - x^{k-1}\|^2 - \|Ax - Ax^{k-1}\|^2, \tag{5.7}$$

to obtain

$$0 = A^T(Ax - b) + \frac{1}{\gamma}(x - x^{k-1}) - A^T(Ax - Ax^{k-1}), \tag{5.8}$$

so that

$$x^k = x^{k-1} - \gamma A^T(Ax^{k-1} - b). \tag{5.9}$$

This is the iterative step of Landweber's algorithm. The sequence $\{x^k\}$ converges to a minimizer x^* of $f(x)$, and x^* minimizes $\|\hat{x} - x^0\|$ over all $\hat{x}$ that minimize $\|Ax - b\|$.

In [9] this same "trick"was used to obtain an elementary proof of convergence of the forward-backward-splitting algorithm [15].

6 The Quadratic Upper Bound Principle

In [3] the authors introduce the *quadratic upper bound principle* as a method for obtaining a majorizing function in optimization transfer. The objective is to minimize the function $f : \mathbb{R}^J \to \mathbb{R}$. If f is twice continuously differentiable, then, for any x and z, we have, according to the extended Mean Value Theorem,

$$f(x) = f(z) + \langle \nabla f(z), x - z \rangle + \frac{1}{2}(x - z)^T \nabla^2 f(w)(x - z), \tag{6.10}$$

for some w on the line segment connecting x and z. If there is a positive-definite matrix B such that $B - \nabla^2 f(w)$ is positive-definite for all w, then we have

$$f(x) \leq f(z) + \langle \nabla f(z), x - z \rangle + \frac{1}{2}(x - z)^T B(x - z). \tag{6.11}$$

Then we have $g(x|z) \geq f(x)$, for all x and z, where

$$g(x|z) \doteq f(z) + \langle \nabla f(z), x - z \rangle + \frac{1}{2}(x - z)^T B(x - z). \tag{6.12}$$

The iterative step is now to minimize $g(x|x^{k-1})$ to get x^k.

The iterative step is equivalent to minimizing

$$G_k(x) = f(x) + \frac{1}{2}(x - x^{k-1})^T B(x - x^{k-1}) - D_f(x, x^{k-1}), \tag{6.13}$$

which is quite similar to the "trick"introduced previously. However, it is not precisely the same, since the authors of [3] do not assume that f is convex, so this is not a particular case of PMAB. Unless f is convex, we cannot assert that this iteration is in the SUMMA class, so we cannot be sure that the iteration reduces $\{f(x^k)\}$ to the infimal value β. This approach also relies on the extended mean value theorem, while our "trick" permits us considerable freeedom in the selection of the function g.

7 Alternating Minimization (AM)

In this section we review the basics of alternating minimization (AM) [16], and then show that AM, PMA and MM are equivalent. Alternating minimization plays an important role in the application of the EM algorithm [18] to medical image reconstruction [28, 29, 6].

7.1 The AM Method

Let $\Phi : X \times Y \to \mathbb{R}_+$, where X and Y are arbitrary nonempty sets. The objective is to find $\hat{x} \in X$ and $\hat{y} \in Y$ such that

$$\Phi(\hat{x}, \hat{y}) \leq \Phi(x, y),$$

for all $x \in X$ and $y \in Y$.

The alternating minimization method [16] is to minimize $\Phi(x, y^{k-1})$ to get x^{k-1} and then to minimize $\Phi(x^{k-1}, y)$ to get y^k. Clearly, the sequence $\{\Phi(x^{k-1}, y^k)\}$ is decreasing and converges to some $\beta^* \geq -\infty$. We want $\beta^* = \Phi(\hat{x}, \hat{y})$, or, at least, for $\beta^* = \beta$, where $\beta = \inf_{x,y} \Phi(x, y)$.

In AM we find x^k by minimizing $\Phi(x, y^k) = \Phi(x, y(x^{k-1}))$. For each x and z in X we define

$$d(x, z) \doteq \Phi(x, y(z)) - \Phi(x, y(x)). \tag{7.1}$$

Clearly, $d(x, z) \geq 0$ and $d(x, x) = 0$, so $d(x, z)$ is a "distance". We obtain x^k by minimizing

$$\Phi(x, y(x^{k-1})) = \Phi(x, y(x)) + \Phi(x, y(x^{k-1})) - \Phi(x, y(x)) = f(x) + d(x, x^{k-1}),$$

which shows that every AM algorithm is also a PMA. Given any AM algorithm, we define $f(x) \doteq \Phi(x, y(x))$. Then the function $g(x|z) \doteq \Phi(x, y(z))$ majorizes $f(x)$. So we see, once again, that AM, PMA and MM are equivalent methods. Now we can obtain conditions on MM algorithms sufficient for $\beta^* = \beta$ from analogous conditions expressed in the language of AM or PMA.

7.2 The Three-Point Property

The *three-point property* (3PP) in [16] is the following: for all $x \in X$ and $y \in Y$ and for all k we have

$$\Phi(x, y^k) - \Phi(x^k, y^k) \geq d(x, x^k). \tag{7.2}$$

The 3PP implies that the AM algorithm, expressed as a PMA, is in the SUMMA class and so is sufficient to have $\beta^* = \beta$.

7.3 The Weak Three-Point Property

The 3PP is stronger than we need to get $\beta^* = \beta$; the weak 3PP implies that the AM algorithm, expressed as a PMA, is in the SUMMA2 class, and so is sufficient for $\beta^* = \beta$. The *weak three-point property* (w3PP) is the following: for all $x \in X$ and $y \in Y$ and for all k we have

$$\Phi(x, y^k) - \Phi(x^k, y^{k+1}) \geq d(x, x^k). \tag{7.3}$$

7.4 Consequences of the w3PP

From the w3PP we find that, for all x and y,

$$d(x, x^{k-1}) - d(x, x^k) \geq \Phi(x^k, y^{k+1}) - \Phi(x, y(x)). \tag{7.4}$$

Since

$$\Phi(x^k, y^{k+1}) - \Phi(x, y(x)) = f(x^k) - f(x)$$

we conclude that, whenever the w3PP holds, we have

$$d(x, x^{k-1}) + f(x) \geq d(x, x^k) + f(x^k), \tag{7.5}$$

for all $x \in X$. This means that AM with the w3PP is in the SUMMA2 class of iterative algorithms, from which it follows that $\beta^* = \beta$.

7.5 When Do We Have $\beta^* = \beta$?

As we have noted, an AM method for which the w3PP holds is in the SUMMA2 class, so that $\beta^* = \beta$. We can formulate this in the language of MM as follows:

$$g(x|x^{k-1}) - g(x|x^k) \geq f(x^k) - f(x) \tag{7.6}$$

for all x. In the language of PMA it becomes

$$d(x, x^{k-1}) - d(x, x^k) \geq f(x^k) - f(x) \tag{7.7}$$

for all x.

We know that all PMAB algorithms are in the SUMMA class. Since PMA is equivalent to MM, this tells us that all MM algorithms for which $g(x|z) - f(x)$ is a Bregman distance will have $\beta^* = \beta$. As we shall see in the next section, the Auslender–Teboulle theory allows us to generalize this result.

8 The Auslender–Teboulle Theory

In [2] Auslender and Teboulle consider proximal minimization algorithms. They show that, if the distance d has associated with it what they call "an induced proximal distance" $h(x, z)$, then $\beta^* = \beta$. It can be shown that, whenever there is an induced proximal distance, then, for any x, we have

$$h(x, x^k) - h(x, x^{k+1}) \geq f(x^k) - f(x) \geq 0. \tag{8.8}$$

Consequently, the algorithm falls into the SUMMA2 class, for which $\beta^* = \beta$ is always true.

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

$$d(x, z) = d_\phi(x, z) \doteq \sum_{j=1}^{J} z_j \phi(\frac{x_j}{z_j}),$$

for functions ϕ having certain properties to be discussed below. In such cases the induced proximal distance is $h(x, z) = \phi''(1)KL(x, z)$, where $KL(x, z)$ is the Kullback–Leibler distance,

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

Then we have

$$\phi''(1) \left(KL(\hat{x}, x^k) - KL(\hat{x}, x^{k+1}) \right) \geq f(x^k) - f(\hat{x}). \tag{8.9}$$

The Hellinger distance,

$$H(x, z) = \sum_{j=1}^{J} (\sqrt{x_j} - \sqrt{z_j})^2,$$

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

$$\phi''(1) \left(1 - \frac{1}{t} \right) \leq \phi'(t) \leq \phi''(1) \log(t). \tag{8.10}$$

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

$$KL(\hat{x}, x^k) - KL(\hat{x}, x^{k+1}) \geq 2 \left(f(x^k) - f(\hat{x}) \right). \tag{8.11}$$

We have already seen that MM algorithms for which $g(x|z) - f(x)$ is a Bregman distance have $\beta^* = \beta$. From [2] we learn that $\beta^* = \beta$ whenever $g(x|z) - f(x) = d_\phi(x, z)$ for functions ϕ satisfying the conditions given above.

9 AM with the Euclidean Distance

9.1 Definitions

In this section we illustrate the use of AM to derive an iterative algorithm to minimize the function $f(x) = \|b - Ax\|^2$, where A is an I by J real matrix and b an I by 1 real vector. Let R be the set of all I by J arrays r with entries $r_{i,j}$ such that $\sum_{j=1}^{J} r_{i,j} = b_i$, for each i. Let Q be the set of all I by J arrays of the form $q(x)$, where $q(x)_{i,j} = A_{i,j}x_j$. For any vectors u and v with the same size, define

$$E(u, v) = \sum_n (u_n - v_n)^2. \tag{9.1}$$

9.2 Pythagorean Identities

We begin by minimizing $E(r, q(x))$ over all $r \in R$. We have the following proposition.

Proposition 9.1 *For all x and r we have*

$$E(r, q(x)) = E(r(x), q(x)) + E(r, r(x)), \tag{9.2}$$

where

$$r(x)_{i,j} = A_{i,j}x_j + \frac{1}{J}(b_i - Ax_i). \tag{9.3}$$

Therefore, $r = r(x)$ is the minimizer of $E(r, q(x))$.

Now we minimize $E(r(x), q(z))$ over z. We have the following proposition.

Proposition 9.2 *For all x and z we have*

$$E(r(x), q(z)) = E(r(x), q(Lx)) + \sum_{j=1}^{J} c_j(Lx_j - z_j)^2, \tag{9.4}$$

where $c_j = \sum_{i=1}^{I} A_{i,j}^2$ and

$$(Lx)_j = Lx_j \doteq x_j + \frac{1}{Jc_j}\sum_{i=1}^{I} A_{i,j}(b_i - Ax_i). \tag{9.5}$$

We omit the proofs of these propositions, which are not deep, but involve messy calculations. Note that

$$\|b - Ax\|^2 = f(x) = JE(r(x), q(x)). \tag{9.6}$$

9.3 The AM Iteration

The iterative step of the algorithm is then

$$x_j^k = Lx_j^{k-1} = x_j^{k-1} + \frac{1}{Jc_j}\sum_{i=1}^{I} A_{i,j}(b_i - Ax_i^{k-1}). \tag{9.7}$$

Applying (9.2) and (9.4) we obtain

$$f(x^{k-1}) = JE(r(x^{k-1}), q(x^{k-1})) = JE(r(x^{k-1}), q(x^k)) + J\sum_{j=1}^{J} c_j(x_j^k - x_j^{k-1})^2$$

$$= JE(r(x^k), q(x^k)) + JE(r(x^{k-1}), r(x^k)) + J\sum_{j=1}^{J} c_j(x_j^k - x_j^{k-1})^2$$

$$= f(x^k) + JE(r(x^{k-1}), r(x^k)) + J\sum_{j=1}^{J} c_j(x_j^k - x_j^{k-1})^2.$$

Therefore,

$$f(x^{k-1}) - f(x^k) = JE(r(x^{k-1}), r(x^k)) + J\sum_{j=1}^{J} c_j(x_j^k - x_j^{k-1})^2 \geq 0,$$

or

$$f(x^{k-1}) - f(x^k) \geq J\sum_{j=1}^{J} c_j(x_j^k - x_j^{k-1})^2 \geq 0, \tag{9.8}$$

from which it follows that the sequence $\{f(x^k)\}$ is decreasing and the sequence $\{\sum_{j=1}^{J} c_j(x_j^k - x_j^{k-1})^2\}$ converges to zero.

The inequality in (9.8) is the *First Monotonicity Property* for the Euclidean case. Since the sequence $\{E(b, Ax^k)\}$ is decreasing, the sequences $\{Ax^k\}$ and $\{x^k\}$ are bounded; let x^* be a cluster point of the sequence $\{x^k\}$. Since the sequence $\{\sum_{j=1}^{J} c_j(x_j^k - x_j^{k-1})^2\}$ converges to zero, it follows that $x^* = Lx^*$.

9.4 Useful Lemmas

We now present several useful lemmas.

Lemma 9.1 *For all x and z we have*

$$E(r(x), r(z)) = \sum_{j=1}^{J} c_j(x_j - z_j)^2 - \frac{1}{J}\sum_{i=1}^{I}(Ax_i - Az_i)^2. \tag{9.9}$$

Lemma 9.2 *For all x and z we have*

$$\frac{1}{J}\sum_{i=1}^{I}(Ax_i - Az_i)^2 \geq \frac{1}{J^2}\sum_{j=1}^{J}\frac{1}{c_j}\left(\sum_{i=1}^{I}A_{i,j}(Ax_i - Az_i)\right)^2. \tag{9.10}$$

Proof: Use Cauchy's Inequality. ∎

Lemma 9.3 *For all x and z we have*

$$E(r(x), r(z)) \geq \sum_{j=1}^{J} c_j(Lx_j - Lz_j)^2. \tag{9.11}$$

It follows from these lemmas that this iterative algorithm is in the SUMMA2 class; for any x we have

$$J\sum_{j=1}^{J} c_j(Lx_j - x_j^k)^2 - J\sum_{j=1}^{J} c_j(Lx_j - x_j^{k+1})^2$$

$$\geq f(x^k) - f(x) + J\sum_{j=1}^{J} c_j(Lx_j - x_j)^2. \tag{9.12}$$

Consequently, the sequence $f(x^k)\}$ converges to the minimum of the function $f(x)$, which must then be $f(x^*)$, and $\{x^k\}$ must converge to x^*.

9.5 Characterizing the Limit

The following proposition characterizes the limit x^*.

Proposition 9.3 *The choice of $\hat{x} = x^*$ minimizes the distance $\sum_{j=1}^{J} c_j(\hat{x}_j - x_j^0)^2$ over all minimizers $\hat{x}$ of $f(x) = \|b - Ax\|^2$.*

Proof: Let $\hat{x}$ be an arbitrary minimizer of $f(x)$. Using the Pythagorean identities we find that

$$JE(r(x^k), q(\hat{x})) = f(\hat{x}) + J\sum_{j=1}^{J} c_j(A\hat{x}_i - Ax^k)_i)^2 - \sum_{i=1}^{I}(A\hat{x}_i - Ax_i^k)^2,$$

and

$$JE(r(x^k), q(\hat{x})) = f(x^{k+1}) + JE(r(x^k), r(x^{k+1})) + J\sum_{j=1}^{J} c_j(\hat{x}_j - x_j^{k+1})^2.$$

Therefore,

$$J\sum_{j=1}^{J} c_j(\hat{x}_j - x_j^k)^2 - J\sum_{j=1}^{J} c_j(\hat{x}_j - x_j^{k+1})^2$$

$$= f(x^{k+1}) - f(\hat{x}) + JE(r(x^k), r(x^{k+1})) + \sum_{i=1}^{I}(A\hat{x}_i - Ax_i^k)^2.$$

Note that the right side of the last equation depends only on $A\hat{x}$ and not directly on $\hat{x}$ itself; therefore the same is true of the left side. Now we sum both sides over the index k to find that $\sum_{j=1}^{J} c_j(\hat{x}_j - x_j^0)^2 - \sum_{j=1}^{J} c_j(\hat{x}_j - x_j^*)^2$ does not depend directly on the choice of $\hat{x}$. The assertion of the proposition follows. ∎

9.6 SUMMA for the Euclidean Case

To get x^k we minimize

$$G_k(x) = JE(r(x^{k-1}), q(x)) = JE(r(x), q(x)) + \big(JE(r(x^{k-1}), q(x)) - JE(r(x), q(x))\big)$$

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

where

$$g_k(x) = \big(JE(r(x^{k-1}), q(x)) - JE(r(x), q(x))\big) = JE(r(x^{k-1}), r(x)).$$

From (9.9) we have

$$g_k(x) = J\sum_{j=1}^{J} c_j(x_j^{k-1} - x_j)^2 - \sum_{i=1}^{I}(Ax_i^{k-1} - Ax_i)^2. \tag{9.13}$$

From

$$G_k(x) - G_k(x^k) =$$

$$JE(r(x^{k-1}), q(x)) - JE(r(x^{k-1}), q(x^k)) = J\sum_{j=1}^{J} c_j(x_j^k - x_j)^2, \tag{9.14}$$

we see that

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

for all x, so that the SUMMA Inequality holds in this case. Therefore, we have

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

for all x, and so

$$g_k(\hat{x}) - g_{k+1}(\hat{x}) \geq f(x^k) - f(\hat{x}) \geq f(x^k) - f(x^{k+1}). \tag{9.15}$$

This is the *Second Monotonicity Property* for the Euclidean case.

9.7 Using the Landweber Algorithm

It is of some interest to consider an alternative approach, using the Landweber (LW) algorithm. The iterative step of the LW algorithm is

$$x_j^k = x_j^{k-1} + \gamma \sum_{i=1}^{I} A_{i,j}(b_i - Ax_i^{k-1}), \tag{9.16}$$

where $0 < \gamma < \frac{2}{\rho(A^TA)}$. We define $\beta_j = \frac{1}{Jc_j}$, $B_{i,j} = \sqrt{\beta_j}A_{i,j}$, and $z_j = x_j/\sqrt{\beta_j}$. Then $Bz = Ax$. The LW algorithm, applied to $Bz = b$ and with $\gamma = 1$, is

$$z^k = z^{k-1} + B^T(b - Bz^{k-1}). \tag{9.17}$$

Since the trace of B^TB is one, the choice of $\gamma = 1$ is allowed. It is known that the LW algorithm converges to the minimizer of $\|b - Bz\|$ for which $\|z - z^0\|$ is minimized. Converting back to the original x^k, we find that we get the same iterative sequence that we got using the AM method. Moreover, we find once again that the sequence $\{x^k\}$ converges to the minimizer x^* of $f(x)$ for which the distance $\sum_{j=1}^{J} c_j(\hat{x}_j - x_j^0)^2$ is minimized over all minimizers $\hat{x}$ of $f(x)$.

The Landweber algorithm applied to the original problem of minimizing $f(x) = \|Ax - b\|^2$ has the iterative step

$$x^k = x^{k-1} - \gamma A^T(Ax^{k-1} - b), \tag{9.18}$$

where $0 < \gamma < \frac{2}{\rho(A^TA)}$. The sequence $\{x^k\}$ converges to the minimizer x^* of $f(x)$ that minimizes $\|\hat{x} - x^0\|$ over all minimizers $\hat{x}$ of $f(x)$.

10 The SMART

In this section we discuss the *simultaneous multiplicative algebraic reconstruction technique* (SMART) [17, 27, 11, 4, 5, 6]. A key step in the proof of convergence is showing that the SMART is in the SUMMA class.

10.1 The Kullback–Leibler or Cross-Entropy Distance

The Kullback–Leibler distance is quite useful in the discussions that follow. For positive numbers s and t, the Kullback–Leibler distance from s to t is

$$KL(s,t) = s\log\frac{s}{t} + t - s. \tag{10.1}$$

Since, for $x > 0$ we have

$$x - 1 - \log x \geq 0$$

and equal to zero if and only if $x = 1$, it follows that

$$KL(s,t) \geq 0,$$

and $KL(s,s) = 0$. We use limits to define $KL(0,t) = t$ and $KL(s,0) = +\infty$. Now we extend the KL distance to nonnegative vectors component-wise. The following lemma is easy to prove.

Lemma 10.1 *For any nonnegative vectors x and z, with $z_+ = \sum_{j=1}^J z_j > 0$, we have*

$$KL(x,z) = KL(x_+, z_+) + KL(x, \frac{x_+}{z_+}z). \tag{10.2}$$

We can extend the KL distance in the obvious way to infinite sequences with nonnegative terms, as well as to nonnegative functions of continuous variables.

10.2 The Problem to be Solved

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 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.

10.3 The SMART Iteration

The SMART algorithm [17, 27, 11, 4, 6] 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

$$x_j^k = x_j^{k-1} \exp\Big(\sum_{i=1}^I P_{ij} \log(y_i/(Px^{k-1})_i)\Big). \tag{10.3}$$

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

$$(Sx)_j \doteq x_j \exp\Big(\sum_{i=1}^I P_{ij} \log(y_i/(Px)_i)\Big). \tag{10.4}$$

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.

10.4 The SMART as AM

In [4] the SMART was derived using the following alternating minimization (AM) approach.

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

$$r(x)_{ij} = x_j P_{ij} y_i / (Px)_i, \tag{10.5}$$

and

$$q(x)_{ij} = x_j P_{ij}. \tag{10.6}$$

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

$$G_k(x) = KL(q(x), r(x^{k-1})) = \sum_{i=1}^{I} \sum_{j=1}^{J} KL(q(x)_{ij}, r(x^{k-1})_{ij}) \tag{10.7}$$

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

$$KL(q(x), r(z)) = KL(q(x), r(x)) + KL(x, z) - KL(Px, Pz); \tag{10.8}$$

and

$$KL(q(x), r(z)) = KL(q(Sz), r(z)) + KL(x, Sz). \tag{10.9}$$

Note that it follows from Equation (10.2) that $KL(x, z) - KL(Px, Pz) \geq 0$.

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

$$G_k(x) = KL(Px, y) + KL(x, x^{k-1}) - KL(Px, Px^{k-1}), \tag{10.10}$$

so that SMART is an AF algorithm and

$$g_k(x) = KL(x, x^{k-1}) - KL(Px, Px^{k-1}). \tag{10.11}$$

Consequently, the sequence $\{f(x^k)\}$ is decreasing and the sequences $\{Px^k\}$ and $\{x^k\}$ are bounded. From

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

we conclude that the SMART is in the SUMMA class. It follows from our discussion of the SUMMA Inequality that, for all $x \geq 0$,

$$g_k(x) + f(x) \geq g_{k+1}(x) + f(x^k). \tag{10.12}$$

Since

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

we see once again that the sequence $\{x^k\}$ is bounded and therefore has a cluster point, x^*, with $f(x^k) \geq f(x^*)$ for all k and $Sx^* = x^*$.

10.5 MM in SMART

At each step of the SMART we minimize the function $KL(q(x), r(x^{k-1}))$ to get x^k. From

$$KL(q(x), r(z)) = KL(q(x), r(x)) + KL(x, z) - KL(Px, Pz) \geq KL(Px, y) \quad (10.13)$$

we see that the function $KL(q(x), r(z)) = g(x|z)$ is a majorizing function for the function $f(x) = KL(Px, y)$.

10.5.1 The First Monotonicity Property for SMART

Using the Pythagorean identities we have

$$KL(Px^k, y) - KL(Px^{k+1}, y) \geq KL(x^k, x^{k+1}). \quad (10.14)$$

10.5.2 The Second Monotonicity Property for SMART

Let $\hat{x}$ be any minimizer of $KL(Px, y)$. We then have

$$\begin{aligned} KL(\hat{x}, x^k) - KL(\hat{x}, x^{k+1}) = KL(Px^{k+1}, y) - KL(P\hat{x}, y) + \\ KL(P\hat{x}, Px^k) + KL(x^{k+1}, x^k) - KL(Px^{k+1}, Px^k) \geq 0. \end{aligned} \quad (10.15)$$

In fact, there is a somewhat more general version of (10.15), that tells us that, since $Sx^* = x^*$ and $f(x^k) \geq f(x^*)$, we can replace $\hat{x}$ with x^* in (10.15), to get

$$\begin{aligned} KL(x^*, x^k) - KL(x^*, x^{k+1}) = KL(Px^{k+1}, y) - KL(Px^*, y) + \\ KL(Px^*, Px^k) + KL(x^{k+1}, x^k) - KL(Px^{k+1}, Px^k) \geq 0. \end{aligned} \quad (10.16)$$

From (10.16) it follows that 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 $\{KL(x^*, x^k)\}$ converges to zero, it follows that $\{x^k\}$ converges to x^*.

10.6 Characterizing the Limit of SMART

Let $\hat{x}$ be any minimizer of $KL(Px, y)$. From Equation (10.15) 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 $y = Px$.

The following theorem summarizes the situation with regard to the SMART [4, 5, 6].

Theorem 10.1 *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.*

11 The EMML Algorithm

In this section we discuss the EMML algorithm [28, 29, 4, 5, 6]. A key step in the proof of convergence is showing that the EMML algorithm is in the SUMMA2 class.

11.1 The EMML Iteration

Once again, we want to find a nonnegative solution or approximate solution x for the linear system of equations $y = Px$. The EMML algorithm will minimize $KL(y, Px)$.

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

$$x_j^k = x_j^{k-1}\Big(\sum_{i=1}^{I} P_{ij}(y_i/(Px^{k-1})_i)\Big). \tag{11.1}$$

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

$$(Mx)_j \doteq x_j\Big(\sum_{i=1}^{I} P_{ij}(y_i/(Px)_i)\Big). \tag{11.2}$$

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.

11.2 The EMML as AM

Now we want to minimize $f(x) = KL(y, Px)$. We have the following helpful *Pythagorean identities*:

$$KL(r(x), q(z)) = KL(r(z), q(z)) + KL(r(x), r(z)); \tag{11.3}$$

and

$$KL(r(x), q(z)) = KL(r(x), q(Mx)) + KL(Mx, z). \tag{11.4}$$

Using these Pythagorean identities we see that, for $\{x^k\}$ given by Equation (11.1), 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. It follows that the EMML sequence $\{x^k\}$ is bounded. In fact, we have

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

Using (10.2) we obtain the following useful inequality:

$$KL(r(x), r(z)) \geq KL(Mx, Mz). \tag{11.5}$$

From

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

and

$$KL(r(x), q(x^k)) = KL(r(x), q(Mx)) + KL(Mx, x^k) = f(x) - KL(Mx, x) + KL(Mx, x^k)$$

we have

$$KL(Mx, x^k) - KL(Mx, x^{k+1}) \geq f(x^k) - f(x) + KL(Mx, x). \tag{11.6}$$

Note that we have used (11.5) here. Therefore, the EMML is in the SUMMA2 class. With x^* a cluster point, we have

$$KL(Mx^*, x^k) - KL(Mx^*, x^{k+1}) \geq f(x^k) - f(x^*) \geq 0. \tag{11.7}$$

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 $Mx^* = x^*$.

The following theorem summarizes the situation with regard to the EMML algorithm [4, 5, 6].

Theorem 11.1 *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 with the SMART, we have been unable to characterize the limit in terms of the starting vector x^0.

11.3 MM in EMML

At each step of the EMML algorithm we minimize $KL(r(x^{k-1}), q(x))$ to get x^k. From

$$KL(r(z), q(x)) = KL(r(x), q(x)) + KL(r(z), r(x)) \tag{11.8}$$

we see that the function

$$KL(r(z), q(x)) = g(x|z) \tag{11.9}$$

is a majorizing function for $f(x) = KL(y, Px)$.

11.4 The First Monotonicity Property for EMML

From the Pythagorean identities we have

$$KL(y, Px^k) - KL(y, Px^{k+1}) = KL(r(x^k), r(x^{k+1})) + KL(x^{k+1}, x^k), \tag{11.10}$$

so that

$$KL(y, Px^k) - KL(y, Px^{k+1}) \geq KL(x^{k+1}, x^k). \tag{11.11}$$

The inequality in (11.11) is called the *First Monotonicity Property* in [19].

11.5 The Second Monotonicity Property for EMML

Let $\hat{x}$ be a minimizer of $f(x) = KL(y, Px)$. Inserting $x = \hat{x}$ into Equation (11.6), we obtain

$$KL(\hat{x}, x^k) - KL(\hat{x}, x^{k+1}) \geq KL(y, Px^k) - KL(y, Px^{k+1}). \tag{11.12}$$

The inequality in (11.12) is called the *Second Monotonicity Property* in [19].

12 The Hellinger Distance

In [19] the authors consider extending the results concerning the KL distance to the Hellinger distance. In particular, they explore the use of AM and MM.

12.1 The Definition of $H(s,t)$

For $s > 0$ and $t > 0$ the Hellinger distance from s to t is

$$H(s,t) = (\sqrt{s} - \sqrt{t})^2. \tag{12.1}$$

As in the case of the KL distance, we can extend H to nonnegative vectors componentwise. In this section we consider the problem of minimizing $H(y, Px)$ given by

$$H(y, Px) = \sum_{i=1}^{I} (\sqrt{y_i} - \sqrt{(Px)_i})^2. \tag{12.2}$$

As in the KL case, we assume that $s_j = \sum_{i=1}^{I} P_{i,j} = 1$ for each j.

12.2 An Alternating-Minimization Approach

In (4.2) of [19] the authors present a majorizing function to be used to generate the iterative sequence. We can show that their majorizing function is $g(x|z) = H(r(z), q(x))$, with the same notation as in the KL case. The following proposition is essentially what appears in [19]. The proof here is simpler than in [19].

Proposition 12.1 *For all $x \geq 0$ and $z \geq 0$ we have*

$$\sum_{j=1}^{J} P_{i,j} \sqrt{x_j z_j} \leq \sqrt{(Px)_i (Pz)_i}. \tag{12.3}$$

Proof: We have

$$\sum_{j=1}^{J} P_{i,j} \sqrt{x_j} \sqrt{z_j} = \sum_{j=1}^{J} \sqrt{P_{i,j} x_j} \sqrt{P_{i,j} z_j}$$

$$\leq \sqrt{\sum_{j=1}^{J} P_{i,j} x_j} \sqrt{\sum_{j=1}^{J} P_{i,j} z_j} = \sqrt{(Px)_i (Pz)_i},$$

by the Cauchy Inequality. ∎

Corollary 12.1 *The function $g(x|z) = H(r(z), q(x))$ majorizes $f(x) = H(y, Px)$.*

Proof: We only need to show that

$$\sum_{i=1}^{I}\sum_{j=1}^{J}\sqrt{r(z)_{i,j}q(x)_{i,j}} \geq \sum_{i=1}^{I}\sqrt{(Px)_i y_i}.$$

We have

$$\sqrt{r(z)_{i,j}q(x)_{i,j}} = P_{i,j}\sqrt{z_j x_j}\sqrt{\frac{y_i}{(Pz)_i}},$$

from which it follows that

$$\begin{aligned}\sum_{i=1}^{I}\sum_{j=1}^{J}\sqrt{r(z)_{i,j}q(x)_{i,j}} &= \sum_{i=1}^{I}\sum_{j=1}^{J}P_{i,j}\sqrt{z_j x_j}\sqrt{\frac{y_i}{(Pz)_i}} \\ &= \sum_{i=1}^{I}\left(\sum_{j=1}^{J}P_{i,j}\sqrt{z_j x_j}\right)\sqrt{\frac{y_i}{(Pz)_i}} \\ \leq \sum_{i=1}^{I}\left(\sqrt{(Pz)_i(Px)_i}\right)\sqrt{\frac{y_i}{(Pz)_i}} &= \sum_{i=1}^{I}\sqrt{(Px)_i y_i}.\end{aligned}$$

∎

Note that Corollary 12.1 can also be obtained by using Lagrange multipliers to minimize $H(r, q(x))$ over all $r = \{r_{i,j}\}$ with $\sum_{j=1}^{J} r_{i,j} = y_i$, for all i.

Corollary 12.2 *For all $x \geq 0$ and $z \geq 0$ we have*

$$\sum_{i=1}^{I}\sqrt{(Px)_i(Pz)_i} \geq \sum_{j=1}^{J}\sqrt{x_j z_j}. \tag{12.4}$$

Proof: From

$$\sum_{j=1}^{J}P_{i,j}\sqrt{x_j z_j} \leq \sqrt{(Px)_i(Pz)_i}$$

we have

$$\sum_{i=1}^{I}\sum_{j=1}^{J}P_{i,j}\sqrt{x_j z_j} \leq \sum_{i=1}^{I}\sqrt{(Px)_i(Pz)_i},$$

so that

$$\sum_{j=1}^{J}\left(\sum_{i=1}^{I}P_{i,j}\right)\sqrt{x_j z_j} \leq \sum_{i=1}^{I}\sqrt{(Px)_i(Pz)_i}.$$

∎

The iterative step of the algorithm is derived by minimizing $H(r(x^{k-1}), q(x))$ to get x^k, with

$$x_j^k = x_j^{k-1}\left(\sum_{i=1}^{I}P_{i,j}\frac{\sqrt{y_i}}{\sqrt{(Px^{k-1})_i}}\right)^2. \tag{12.5}$$

We can write $x^k = Tx^{k-1}$, where T is the operator

$$Tx_j = x_j \left(\sum_{i=1}^{I} P_{i,j} \frac{\sqrt{y_i}}{\sqrt{(Px)_i}} \right)^2 . \tag{12.6}$$

Since $g(x|z)$ majorizes $f(x)$, it follows easily that the sequence $\{f(x^k)\}$ is decreasing, so that the sequences $\{Px^k\}$ and $\{x^k\}$ are bounded.

In the EMML case we saw that

$$\sum_{i=1}^{I} y_i = \sum_{j=1}^{J} Mx_j,$$

while for SMART we have

$$\sum_{i=1}^{I} y_i \geq \sum_{j=1}^{J} Sx_j.$$

In the Hellinger case we shall see that

$$\sum_{i=1}^{I} y_i \geq \sum_{j=1}^{J} Tx_j.$$

In the EMML case we have the Pythagorean identity

$$KL(r(x), q(z)) = KL(r(x), q(Mx)) + KL(Mx, z), \tag{12.7}$$

while in the Hellinger case we have the analogous Pythagorean identity

$$H(r(x), q(z)) = H(r(x), q(Tx)) + H(Tx, z), \tag{12.8}$$

so that

$$H(r(x), q(x)) = H(r(x), q(Tx)) + H(Tx, x). \tag{12.9}$$

We note that, unlike the KL distance, the Hellinger distance is symmetric; we have

$$H(x, z) = H(z, x). \tag{12.10}$$

Lemma 12.1 *For every $x \geq 0$ we have*

$$H(r(x), q(Tx)) = \sum_{i=1}^{I} y_i - \sum_{j=1}^{J} (Tx)_j \geq 0, \tag{12.11}$$

so that the set $\{Tx | x \geq 0\}$ is bounded.

Since minimizing $f(x) = H(r(x), q(x))$ is equivalent to minimizing $H(r(x), q(Tx))$, it follows that minimizing $f(x)$ is equivalent to maximizing $\sum_{j=1}^{J} Tx_j$. In the EMML case we have

$$f(x) = KL(y, Px) = KL(r(x), q(x)), \tag{12.12}$$

while in the Hellinger case we have the analogous result

$$f(x) = H(y, Px) = H(r(x), q(x)). \tag{12.13}$$

In the EMML case we use the other Pythagorean identity

$$KL(r(z), q(x)) = KL(r(x), q(x)) + KL(r(z), r(x)) \tag{12.14}$$

to show that

$$KL(r(z), q(x)) = g(x|z) \tag{12.15}$$

is a majorizing function for $f(x) = KL(y, Px)$. In the Hellinger case we have shown that

$$H(r(z), q(x)) \geq H(r(x), q(x)), \tag{12.16}$$

for all $x \geq 0$. It would be nice if we had an analogue of Equation (12.14) for the Hellinger case. Said another way, can we find a simple expression for

$$H(r(z), q(x)) - H(r(x), q(x))?$$

By analogy with the EMML case, we might expect to have

$$H(r(z), q(x)) - H(r(x), q(x)) = H(r(z), r(x)). \tag{12.17}$$

Actually, we don't need this much; it would be enough to prove that Equation (12.17) holds for $x = Tz$.

It is worth noting here that perhaps we should consider analogies, not just with the EMML, but with the SMART also. The Hellinger distance is symmetric, so that $H(y, Px) = H(Px, y)$, whereas $KL(y, Px)$ and $KL(Px, y)$ are not the same. In the SMART case we have the inequality

$$KL(x, z) \geq KL(Px, Pz). \tag{12.18}$$

This holds as well for the Hellinger distance.

Lemma 12.2 *For all $x \geq 0$ and $z \geq 0$ we have*

$$H(x,z) \geq H(Px,Pz). \tag{12.19}$$

Proof: We have

$$\begin{aligned} H(Px,Pz) &= \sum_{i=1}^{I} \left((Px)_i + (Pz)_i - 2\sqrt{(Px)_i (Pz)_i} \right) \\ &= \sum_{j=1}^{J} (x_j + z_j) - 2 \sum_{i=1}^{I} \sqrt{(Px)_i (Pz)_i} \\ &\leq \sum_{j=1}^{J} (x_j + z_j) - 2 \sum_{j=1}^{J} \sqrt{x_j z_j}, \end{aligned}$$

by (12.4). ∎

12.3 Convergence

From the discussion above we have

$$H(y, Px^k) - H(y, Px^{k+1}) \geq H(x^k, x^{k+1}), \tag{12.20}$$

so that the sequence $\{H(y, Px^k)\}$ is decreasing and the sequence $\{H(x^k, x^{k+1})\}$ converges to zero. Since the sequence $\{x^k\}$ is bounded, it has a cluster point, call it $\hat{x}$ which must then be a fixed point of T. The sequence $\{H(y, Px^k)\}$ then converges to $H(y, P\hat{x})$. In [19] it was shown that, if $\hat{x}$ minimizes $f(x)$, then

$$KL(\hat{x}, x^k) - KL(\hat{x}, x^{k+1}) \geq 2 \left(H(y, Px^{k+1}) - H(y, P\hat{x}) \right). \tag{12.21}$$

It follows that the sequence $\{x^k\}$ converges to $\hat{x}$.

13 Pearson's ϕ^2 Distance

In [19] the authors consider extending the results concerning the KL and Hellinger distances to the ϕ^2-distance of Pearson . In particular, they explore the use of AM and MM.

13.1 The Definition of $\phi^2(s,t)$

For $s > 0$ and $t > 0$ Pearson's ϕ^2 distance from s to t is

$$\phi^2(s,t) = \frac{(s-t)^2}{t} \tag{13.1}$$

As in the cases of the KL and H distances, we can extend ϕ^2 to nonnegative vectors component-wise. Note that $\phi^2(s,t)$ is not symmetric. In this section we consider the problem of minimizing $\phi^2(y, Px)$ given by

$$\phi^2(y, Px) = \sum_{i=1}^{I} \frac{(y_i - (Px)_i)^2}{(Px)_i}. \tag{13.2}$$

As in the previous cases, we assume that $s_j = 1$ for each j.

13.2 An Alternating-Minimization Approach

In (5.4) of [19] the authors present a majorizing function to be used to generate the iterative sequence. We can show that their majorizing function is $g(x|z) = \phi^2(r(z), q(x))$, with the same notation as in the KL and H cases. The following proposition is essentially what appears in [19]; the proof given here is simpler, however.

Proposition 13.1 *For all $x > 0$ and $z > 0$ we have*

$$\sum_{j=1}^{J} P_{i,j} \frac{x_j^2}{z_j} \geq \frac{(Px)_i^2}{(Pz)_i}. \tag{13.3}$$

Proof: We have

$$\begin{aligned}(Px)_i = \sum_{j=1}^{J} P_{i,j} x_j &= \sum_{j=1}^{J} \sqrt{P_{i,j} z_j} \sqrt{P_{i,j} \frac{x_j^2}{z_j}} \\ &\leq \sqrt{\sum_{j=1}^{J} P_{i,j} z_j} \sqrt{\sum_{j=1}^{J} P_{i,j} \frac{x_j^2}{z_j}},\end{aligned}$$

so that

$$(Px)_i^2 \leq (Pz)_i \sum_{j=1}^{J} P_{i,j} \frac{x_j^2}{z_j}.$$

∎

Corollary 13.1 *For all $x > 0$ and $z > 0$ we have*

$$\sum_{i=1}^{I} \frac{(Px)_i^2}{(Pz)_i} \leq \sum_{j=1}^{J} \frac{x_j^2}{z_j}. \tag{13.4}$$

Corollary 13.2 *The function* $g(x|z) = \phi^2(r(z), q(x))$ *majorizes* $\phi^2(y, Px)$.

Note that Corollary 13.2 can also be obtained by using Lagrange multipliers to minimize $\phi^2(r, q(x))$ over all $r = \{r_{i,j}\}$ with $\sum_{j=1}^{J} r_{i,j} = y_i$, for all i.

Corollary 13.3 *For each* $x > 0$ *and* $z > 0$ *we have*

$$\phi^2(x, z) \geq \phi^2(Px, Pz). \tag{13.5}$$

The iterative step of the algorithm is derived by minimizing $\phi^2(r(x^{k-1}), q(x))$ to get x^k given by

$$x_j^k = x_j^{k-1} \sqrt{\sum_{i=1}^{I} P_{i,j} \left(\frac{y_i}{(Px^{k-1})_i} \right)^2}. \tag{13.6}$$

With R the operator defined by

$$(Rx)_j = Rx_j \doteq x_j \sqrt{\sum_{i=1}^{I} P_{i,j} \left(\frac{y_i}{(Px)_i} \right)^2}, \tag{13.7}$$

we can write $x^k = Rx^{k-1}$. An easy calculation shows that $\phi^2(Rz, x) = \phi^2(q(Rz), q(x))$ and

$$\phi^2(r(z), q(x)) = \phi^2(r(z), q(Rz)) + \phi^2(Rz, x). \tag{13.8}$$

Since $g(x|z)$ majorizes $f(x)$ it follows that the sequence $\{f(x^k)\}$ is decreasing, so that the sequences $\{Px^k\}$ and $\{x^k\}$ are bounded. We also have

$$\phi^2(r(x), q(Rx)) = \sum_{j=1}^{J} (Rx)_j - \sum_{i=1}^{I} y_i \geq 0.$$

14 Just a Coincidence?

As we have seen, the KL distance appears, apparently uninvited, in (12.21). In [2] a similar thing happens, as (14.1) shows, prompting us to ask if this is just a coincidence, or if something deeper is going on here.

In proximal minimization algorithms (PMA) we obtain an iterative method for minimizing a function $f(x)$ by minimizing

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

to get the next iterate x^k. Here $d(x,z) \geq 0$ and $d(x,x) = 0$, for all x and z. It follows easily that the sequence $\{f(x^k)\}$ is decreasing to a limit $\beta^* \geq -\infty$. We have discussed what additional restrictions should be placed on the distance d to guarantee that

$$\beta^* = \beta \doteq \inf_x \{f(x)\}.$$

For the Hellinger distance we have $H(x,z) = d_\phi(x,z)$, for $\phi(t) = (\sqrt{t}-1)^2$, so that, according to [2],

$$KL(\hat{x}, x^k) - KL(\hat{x}, x^{k+1}) \geq 2\left(f(x^k) - f(\hat{x})\right). \tag{14.1}$$

This looks a lot like (12.21).

Of course, the problems are not quite the same; in [2] they are trying to minimize some unrelated function $f(x)$, using the Hellinger distance in the PMA framework, while we are trying to minimize $H(y, Px) = H(Px, y)$ using alternating minimization. However, the resemblance between (12.21) and (14.1) must be more than a coincidence, mustn't it?

15 Acknowledgments

The authors thank Professor Paul Eggermont of the University of Delaware and Professor Hung Phan of the University of Massachusetts Lowell for helpful discussions, and Professor Eggermont for making available the preprint [19].

References

1. Ahn, S., Fessler, J., Blatt, D., and Hero, A. (2006) "Convergent incremental optimization transfer algorithms: application to tomography." *IEEE Transactions on Medical Imaging*, **25(3)**, pp. 283–296.

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

3. Böhning, D., and Lindsey, B.G. (1988) "Monotonicity of quadratic approximation algorithms." *Ann Instit Stat Math*, **40**, pp. 641–663.

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

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

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

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

8. Byrne, C. (2013) "Alternating minimization as sequential unconstrained minimization: a survey." *Journal of Optimization Theory and Applications*, electronic **154(3)**, DOI 10.1007/s1090134-2, (2012), and hardcopy **156(3)**, February, 2013, pp. 554–566.

9. Byrne, C. (2014) "An elementary proof of convergence of the forward-backward splitting algorithm." *Journal of Nonlinear and Convex Analysis* , **15(4)**, pp. 681–691.

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

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

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

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

14. Chi, E., Zhou, H., and Lange, K. (2014) "Distance majorization and its applications." *Mathematical Programming*, **146 (1-2)**, pp. 409–436.

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

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

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

18. 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.

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

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

21. Erdogan, H., and Fessler, J. (1999) "Monotonic algorithms for transmission tomography." *IEEE Transactions on Medical Imaging*, **18(9)**, pp. 801–814.

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

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

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

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

26. 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.

27. Schmidlin, P. (1972) "Iterative separation of sections in tomographic scintigrams." *Nuklearmedizin*, **11**, pp. 1–16.

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

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