IOP PUBLISHING INVERSE PROBLEMS

Inverse Problems **24** (2008) 015013 (27pp) doi:10.1088/0266-5611/24/1/015013

Sequential unconstrained minimization algorithms for constrained optimization

Charles Byrne

Department of Mathematical Sciences, University of Massachusetts at Lowell, Lowell, MA 01854, USA

E-mail: Charles_Byrne@uml.edu

Received 11 October 2007, in final form 30 November 2007
Published 9 January 2008
Online at stacks.iop.org/IP/24/015013

Abstract

The problem of minimizing a function $f(x) : R^J \to R$, subject to constraints on the vector variable x, occurs frequently in inverse problems. Even without constraints, finding a minimizer of $f(x)$ may require iterative methods. We consider here a general class of iterative algorithms that find a solution to the constrained minimization problem as the limit of a sequence of vectors, each solving an unconstrained minimization problem. Our sequential unconstrained minimization algorithm (SUMMA) is an iterative procedure for constrained minimization. At the kth step we minimize the function

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

to obtain x^k. The auxiliary functions $g_k(x) : D \subseteq R^J \to R_+$ are nonnegative on the set D, each x^k is assumed to lie within D, and the objective is to minimize the continuous function $f : R^J \to R$ over x in the set $C = \overline{D}$, the closure of D. We assume that such minimizers exist, and denote one such by $\hat{x}$. We assume that the functions $g_k(x)$ satisfy the inequalities

$$0 \leqslant g_k(x) \leqslant G_{k-1}(x) - G_{k-1}(x^{k-1}),$$

for $k = 2, 3, \ldots$. Using this assumption, we show that the sequence $\{f(x^k)\}$ is decreasing and converges to $f(\hat{x})$. If the restriction of $f(x)$ to D has bounded level sets, which happens if $\hat{x}$ is unique and $f(x)$ is closed, proper and convex, then the sequence $\{x^k\}$ is bounded, and $f(x^*) = f(\hat{x})$, for any cluster point x^*. Therefore, if $\hat{x}$ is unique, $x^* = \hat{x}$ and $\{x^k\} \to \hat{x}$. When $\hat{x}$ is not unique, convergence can still be obtained, in particular cases. The SUMMA includes, as particular cases, the well-known barrier- and penalty-function methods, the simultaneous multiplicative algebraic reconstruction technique (SMART), the proximal minimization algorithm of Censor and Zenios, the entropic proximal methods of Teboulle, as well as certain cases of gradient descent and the Newton–Raphson method. The proof techniques used for SUMMA can be extended to obtain related results for the induced proximal distance method of Auslander and Teboulle.

1. Introduction

In many inverse problems, we have measured data pertaining to the object x, which may be, for example, a vectorized image, as well as prior information about x, such that its entries are nonnegative. Tomographic imaging is a good example. We want to find an estimate of x that is (more or less) consistent with the data, as well as conforming to the prior constraints. The measured data and prior information are usually not sufficient to determine a unique x and some form of optimization is performed. For example, we may seek the image x for which the entropy is maximized, or a minimum-norm least-squares solution.

There are many well-known methods for minimizing a function $f : R^J \rightarrow R$; we can use the Newton–Raphson algorithm or any of its several approximations, or nonlinear conjugate-gradient algorithms, such as the Fletcher–Reeves, Polak–Ribiere, or Hestenes–Stiefel methods. When the problem is to minimize the function $f(x)$, subject to constraints on the variable x, the problem becomes much more difficult. For such constrained minimization, we can employ sequential unconstrained minimization algorithms [15].

We assume that $f : R^J \rightarrow (-\infty, +\infty]$ is a continuous function. Our objective is to minimize $f(x)$ over x in some given closed nonempty set C. At the kth step of a sequential unconstrained minimization algorithm we minimize a function $G_k(x)$ to obtain the vector x^k. We shall assume throughout that a global minimizer x^k exists for each k. The existence of these minimizers can be established, once additional conditions, such as convexity, are placed on the functions $G_k(x)$; see, for example, Fiacco and McCormick [15], p 95. We shall consider briefly, near the end of this paper, the issue of computing the x^k.

In the best case, the sequence $\{x^k\}$ converges to a constrained minimizer of the original objective function $f(x)$. Obviously, the functions $G_k(x)$ must involve both the function $f(x)$ and the set C. Those methods for which each x^k is *feasible*, that is, each x^k is in C, are called *interior-point* methods, while those for which only the limit of the sequence is in C are called *exterior-point* methods. Barrier-function algorithms are typically interior-point methods, while penalty-function algorithms are exterior-point methods. The purpose of this paper is to present a fairly broad class of sequential unconstrained minimization algorithms, which we call SUMMA. The SUMMA include both barrier- and penalty-function algorithms, as well as proximity-function methods of Teboulle [24] and Censor and Zenios [13], the simultaneous multiplicative algebraic reconstruction technique (SMART) [6], as well as certain cases of gradient descent and the Newton–Raphson method. The proof techniques used for SUMMA can be extended to obtain related results for the induced proximal distance method of Auslander and Teboulle [3].

The sequential unconstrained minimization algorithms (SUMMA) we present here minimize functions of the form

$$G_k(x) = f(x) + g_k(x), \tag{1.1}$$

to obtain the next iterate x^k, with the auxiliary functions $g_k(x)$ chosen so that

$$0 \leqslant g_{k+1}(x) \leqslant G_k(x) - G_k(x^k), \tag{1.2}$$

for $k = 1, 2, \ldots$. We assume throughout that there exists $\hat{x}$ minimizing the function $f(x)$ over x in C. Our main results are that the sequence $\{f(x^k)\}$ is monotonically decreasing to $f(\hat{x})$, and, subject to certain conditions on the function $f(x)$, the sequence $\{x^k\}$ converges to a feasible x^* with $f(x^*) = f(\hat{x})$.

We begin with a brief review of several types of sequential unconstrained minimization methods, including those mentioned previously. Then we state and prove the convergence results for the SUMMA. Finally, we show that each of these methods reviewed previously is a particular case of the SUMMA.

2. Barrier-function methods (I)

Let $b(x) : R^J \to (-\infty, +\infty]$ be continuous, with effective domain the set

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

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

$$f(x) + \frac{1}{k} b(x) \tag{2.1}$$

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

Barrier functions typically have the property that $b(x) \to +\infty$ as x approaches the boundary of D, so not only is x^k prevented from leaving D, it is discouraged from approaching the boundary.

2.1. Examples of barrier functions

Consider the convex programming (CP) problem of minimizing the convex function $f : R^J \to R$, subject to $g_i(x) \leqslant 0$, where each $g_i : R^J \to R$ is convex, for $i = 1, \ldots, I$. Let $D = \{x | g_i(x) < 0, i = 1, \ldots, I\}$; then D is open. We consider two barrier functions appropriate for this problem.

2.1.1. The logarithmic barrier function. A suitable barrier function is the *logarithmic barrier function*

$$b(x) = -\sum_{i=1}^{I} \log(-g_i(x)). \tag{2.2}$$

The function $-\log(-g_i(x))$ is defined only for those x in D, and is positive for $g_i(x) > -1$. If $g_i(x)$ is near zero, then so is $-g_i(x)$ and $b(x)$ will be large.

2.1.2. The inverse barrier function. Another suitable barrier function is the *inverse barrier function*

$$b(x) = \sum_{i=1}^{I} \frac{-1}{g_i(x)}, \tag{2.3}$$

defined for those x in D.

2.1.3. An illustration. We minimize the function $f(u, v) = u^2 + v^2$, subject to the constraint that $u + v \geqslant 1$. The constraint is then written $g(u, v) = 1 - (u + v) \leqslant 0$. We use the logarithmic barrier. The vector $x^k = (u^k, v^k)$ minimizing the function

$$G_k(x) = u^2 + v^2 - \frac{1}{k} \log(u + v - 1)$$

has entries

$$u^k = v^k = \frac{1}{4} + \frac{1}{4}\sqrt{1 + \frac{4}{k}}.$$

Note that $u^k + v^k > 1$, so each x^k satisfies the constraint. As $k \to +\infty$, x^k converges to $\left(\frac{1}{2}, \frac{1}{2}\right)$, which is the solution to the original problem.

3. Penalty-function methods (I)

When we add a barrier function to $f(x)$ we restrict the domain. When the barrier function is used in a sequential unconstrained minimization algorithm, the vector x^k that minimizes the function $f(x) + \frac{1}{k}b(x)$ lies in the effective domain D of $b(x)$, and we prove that, under certain conditions, the sequence $\{x^k\}$ converges to a minimizer of the function $f(x)$ over the closure of D. The constraint of lying within the set $\overline{D}$ is satisfied at every step of the algorithm; for that reason such algorithms are called interior-point methods. Constraints may also be imposed using a penalty function. In this case, violations of the constraints are discouraged, but not forbidden. When a penalty function is used in a sequential unconstrained minimization algorithm, the x^k need not satisfy the constraints; only the limit vector need be feasible. As we shall see, under conditions to be specified later the penalty-function method can be used to minimize a continuous function $f(x)$ over the nonempty set of minimizers of another continuous function $p(x)$.

A penalty function $p(x)$ is a non-negative function that is positive outside C and zero on C. At the kth step of the algorithm we minimize the function

$$f(x) + kp(x), \tag{3.1}$$

to obtain x^k. Typically, $p(x^k) > 0$ and no x^k lies within C.

3.1. Examples of penalty functions

Consider the CP problem again. We wish to minimize the convex function $f(x)$ over all x for which the convex functions $g_i(x) \leqslant 0$, for $i = 1, \ldots, I$.

3.1.1. The absolute-value penalty function. We let $g_i^+(x) = \max\{g_i(x), 0\}$, and

$$p(x) = \sum_{i=1}^{I} g_i^+(x). \tag{3.2}$$

This is the absolute-value penalty function; it penalizes violations of the constraints $g_i(x) \leqslant 0$, but does not forbid such violations. Then, for $k = 1, 2, \ldots$, we minimize

$$f(x) + kp(x), \tag{3.3}$$

to obtain x^k. As $k \to +\infty$, the penalty function becomes more heavily weighted, so that, in the limit, the constraints $g_i(x) \leqslant 0$ should hold. Because only the limit vector satisfies the constraints, and the x^k are allowed to violate them, such a method is called an *exterior-point* method.

3.1.2. The Courant–Beltrami penalty function. The *Courant–Beltrami* penalty-function method is similar, but uses

$$p(x) = \sum_{i=1}^{I} \left[g_i^+(x)\right]^2. \tag{3.4}$$

3.1.3. The quadratic-loss penalty function. Penalty methods can also be used with equality constraints. Consider the problem of minimizing the convex function $f(x)$, subject to the constraints $g_i(x) = 0, i = 1, \ldots, I$. The *quadratic-loss* penalty function is

$$p(x) = \frac{1}{2}\sum_{i=1}^{I} (g_i(x))^2. \tag{3.5}$$

The inclusion of a penalty term can serve purposes other than to impose constraints on the location of the limit vector. In image processing, it is often desirable to obtain a reconstructed image that is locally smooth, but with well-defined edges. Penalty functions that favor such images can then be used in the iterative reconstruction [16]. We survey several instances in which we would want to use a penalized objective function.

3.1.4. Regularized least-squares. Suppose we want to solve the system of equations $Ax = b$. The problem may have no exact solution, precisely one solution, or there may be infinitely many solutions. If we minimize the function

$$f(x) = \tfrac{1}{2}||Ax - b||_2^2,$$

we get a *least-squares* solution, generally, and an exact solution, whenever exact solutions exist. When the matrix A is ill conditioned, small changes in the vector b can lead to large changes in the solution. When the vector b comes from measured data, the entries of b may include measurement errors, so that an exact solution of $Ax = b$ may be undesirable, even when such exact solutions exist; exact solutions may correspond to x with unacceptably large norm, for example. In such cases, we may, instead, wish to minimize a function such as

$$\frac{1}{2}||Ax - b||_2^2 + \frac{\epsilon}{2}||x - z||_2^2, \tag{3.6}$$

for some vector z. If $z = 0$, the minimizing vector x_ϵ is then a *norm-constrained* least-squares solution. We then say that the least-squares problem has been *regularized*. In the limit, as $\epsilon \to 0$, these regularized solutions x_ϵ converge to the least-squares solution closest to z.

Suppose the system $Ax = b$ has infinitely many exact solutions. Our problem is to select one. Let us select z that incorporates features of the desired solution, to the extent that we know them *a priori*. Then, as $\epsilon \to 0$, the vectors x_ϵ converge to the exact solution closest to z. For example, taking $z = 0$ leads to the *minimum-norm solution*.

3.1.5. Minimizing cross-entropy. In image processing, it is common to encounter systems $Px = y$ in which all the terms are non-negative. In such cases, it may be desirable to solve the system $Px = y$, approximately, perhaps, by minimizing the *cross-entropy* or *Kullback–Leibler distance*

$$\mathrm{KL}(y, Px) = \sum_{i=1}^{I} \left(y_i \log \frac{y_i}{(Px)_i} + (Px)_i - y_i \right), \tag{3.7}$$

over vectors $x \geqslant 0$. When the vector y is noisy, the resulting solution, viewed as an image, can be unacceptable. It is wise, therefore, to add a penalty term, such as $p(x) = \epsilon\,\mathrm{KL}(z, x)$, where $z > 0$ is a prior estimate of the desired x [6, 17, 18, 25].

A similar problem involves minimizing the function $\mathrm{KL}(Px, y)$. Once again, noisy results can be avoided by including a penalty term, such as $p(x) = \epsilon\,\mathrm{KL}(x, z)$ [6].

3.1.6. The Lagrangian in convex programming. When there is a sensitivity vector λ for the CP, minimizing $f(x)$ is equivalent to minimizing the Lagrangian,

$$f(x) + \sum_{i=1}^{I} \lambda_i g_i(x) = f(x) + p(x); \tag{3.8}$$

in this case, the addition of the second term, $p(x)$, serves to incorporate the constraints $g_i(x) \leqslant 0$ in the function to be minimized, turning a constrained minimization problem into an unconstrained one. The problem of minimizing the Lagrangian still remains, though. We may have to solve that problem using an iterative algorithm.

3.1.7. Moreau's proximity-function method. The Moreau envelope [19] of the function f is the function

$$m_f(z) = \inf_x \left\{ f(x) + \tfrac{1}{2}||x - z||_2^2 \right\}, \tag{3.9}$$

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 [22, 23]). In similar fashion, we can define $m_{f^*}z$ and $\text{prox}_{f^*}z$, where $f^*(z)$ denotes the function conjugate to f.

Proposition 3.1. *The infimum of $m_f(z)$, over all z, is the same as the infimum of $f(x)$, over all x.*

Proof. We have

$$\begin{aligned}\inf_z m_f(z) &= \inf_z \inf_x \left\{ f(x) + \tfrac{1}{2}||x - z|_2^2 \right\} \\ &= \inf_x \inf_z \left\{ f(x) + \tfrac{1}{2}||x - z|_2^2 \right\} = \inf_x \left\{ f(x) + \tfrac{1}{2} \inf_z ||x - z||_2^2 \right\} = \inf_x f(x). \qquad \square\end{aligned}$$

The minimizers of $m_f(z)$ and $f(x)$ are the same, as well. Therefore, one way to use Moreau's method is to replace the original problem of minimizing the possibly non-smooth function $f(x)$ with the problem of minimizing the smooth function $m_f(z)$. Another way is to convert Moreau's method into a sequential minimization algorithm, replacing z with x^{k-1} and minimizing with respect to x to obtain x^k. As we shall see, this leads to the proximal minimization algorithm to be discussed below.

3.2. The roles penalty functions play

From the examples just surveyed, we can distinguish several distinct roles that penalty functions can play.

3.2.1. Impose constraints. The first role is to penalize violations of constraints, as part of sequential minimization, or even to turn a constrained minimization into an equivalent unconstrained one: the Absolute-Value and Courant–Beltrami penalty functions penalize violations of the constraints $g_i(x) \leqslant 0$, while quadratic-loss penalty function penalizes violations of the constraints $g_i(x) = 0$. The augmented objective functions $f(x) + kp(x)$ now become part of a sequential unconstrained minimization method. It is sometimes possible for $f(x)$ and $f(x) + p(x)$ to have the same minimizers, or for constrained minimizers of $f(x)$ to be the same as unconstrained minimizers of $f(x) + p(x)$, as happens with the Lagrangian in the CP problem.

3.2.2. Regularization. The second role is regularization: in the least-squares problem, the main purpose for adding the norm-squared penalty function in equation (3.6) is to reduce sensitivity to noise in the entries of the vector b. Also, regularization will usually turn a problem with multiple solutions into one with a unique solution.

3.2.3. Incorporate prior information. The third role is to incorporate prior information: when $Ax = b$ is under-determined, using the penalty function $\epsilon||x - z||_2^2$ and letting $\epsilon \to 0$ encourages the solution to be close to the prior estimate z.

3.2.4. Simplify calculations. A fourth role that penalty functions can play is to simplify calculation: in the case of cross-entropy minimization, adding the penalty functions $\mathrm{KL}(z, x)$ and $\mathrm{KL}(x, z)$ to the objective functions $\mathrm{KL}(y, Px)$ and $\mathrm{KL}(Px, y)$, respectively, regularizes the minimization problem. But, as we shall see later, the SMART algorithm minimizes $\mathrm{KL}(Px, y)$ by using a sequential approach, in which each minimizer x^k can be calculated in closed form.

3.2.5. Sequential unconstrained minimization. More generally, a fifth role for penalty functions is as part of sequential minimization. Here the goal is to replace one computationally difficult minimization with a sequence of simpler ones. Clearly, one reason for the difficulty can be that the original problem is constrained, and the sequential approach uses a series of unconstrained minimizations, penalizing violations of the constraints through the penalty function. However, there are other instances in which the sequential approach serves to simplify the calculations, not to remove constraints, but, perhaps, to replace a non-differentiable objective function with a differentiable one, or a sequence of differentiable ones, as in Moreau's method.

4. Proximity-function minimization (I)

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

$$D_h(x, z) = h(x) - h(z) - \langle \nabla h(z), x - z \rangle. \tag{4.1}$$

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

4.1. Proximal minimization algorithm

At the kth step of the *proximal minimization algorithm* (PMA) [8], we minimize the function

$$G_k(x) = f(x) + D_h(x, x^{k-1}), \tag{4.2}$$

to obtain x^k. The function

$$g_k(x) = D_h(x, x^{k-1}) \tag{4.3}$$

is nonnegative and $g_k(x^{k-1}) = 0$. We assume that each x^k lies in int D.

4.2. The method of Auslander and Teboulle

In [3] Auslander and Teboulle consider an iterative method similar to the PMA, in which, at the kth step, one minimizes the function

$$F_k(x) = f(x) + d(x, x^{k-1}) \tag{4.4}$$

to obtain x^k. Their distance $d(x, y)$ is not assumed to be a Bregman distance. Instead, they assume that the distance d has an associated *induced proximal distance* $H(a, b) \geqslant 0$, finite for a and b in D, with $H(a, a) = 0$ and

$$\langle \nabla_1 d(b, a), c - b \rangle \leqslant H(c, a) - H(c, b), \tag{4.5}$$

for all c in D. The notation $\nabla_1 d(x, y)$ denotes the gradient with respect to the vector variable x.

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

$$\langle \nabla_1 d(b,a), c-b\rangle = D_h(c,a) - D_h(c,b) - D_h(b,a) \tag{4.6}$$

we see that D_h has $H = D_h$ for its associated induced proximal distance, so D_h is *self-proximal*, in the terminology of [3].

5. The simultaneous MART (SMART) (I)

Our next example is the simultaneous multiplicative algebraic reconstruction technique (SMART). For $a > 0$ and $b > 0$, the Kullback–Leibler distance, $\mathrm{KL}(a,b)$, is defined as

$$\mathrm{KL}(a,b) = a\log\frac{a}{b} + b - a. \tag{5.1}$$

In addition, $\mathrm{KL}(0,0) = 0, \mathrm{KL}(a,0 = +\infty$ and $\mathrm{KL}(0,b) = b$. The KL distance is then extended to nonnegative vectors coordinate wise.

5.1. The SMART iteration

The SMART minimizes the function $f(x) = \mathrm{KL}(Px, y)$, over nonnegative vectors x. Here y is a vector with positive entries, and P is a matrix with nonnegative entries, such that $s_j = \sum_{i=1}^I P_{ij} > 0$. For notational convenience, we shall assume that the system $y = Px$ has been normalized so that $s_j = 1$, for each j. Denote by $\mathcal{X}$ the set of all nonnegative x for which the vector Px has only positive entries.

Having found the vector x^{k-1}, the next vector in the SMART sequence is x^k, with entries given by

$$x_j^k = x_j^{k-1}\exp\left(\sum_{i=1}^I P_{ij}\log(y_i/(Px^{k-1})_i)\right). \tag{5.2}$$

5.2. SMART as alternating minimization

In [6] the SMART was derived using the following alternating minimization approach.

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

$$r(x)_{ij} = x_j P_{ij} y_i/(Px)_i, \tag{5.3}$$

and

$$q(x)_{ij} = x_j P_{ij}. \tag{5.4}$$

The iterative step of the SMART is to minimize the function $\mathrm{KL}(q(x), r(x^{k-1}))$ to obtain $x = x^k$. Note that $f(x) = \mathrm{KL}(q(x), r(x))$.

Now we establish the basic results for the SUMMA.

6. Convergence theorems for SUMMA

At the kth step of the SUMMA we minimize the function $G_k(x)$ to obtain x^k. In practice, of course, this minimization may need to be performed iteratively; we shall address this issue only in passing this paper, and shall assume throughout that x^k can be computed. We make the following additional assumptions.

Assumption 1. The functions $g_k(x)$ are finite-valued and continuous on a set D in R^J, with $C = \overline{D}$.

Assumption 2. There is $\hat{x}$ in C with $f(\hat{x}) \leqslant f(x)$, for all x in C.

Assumption 3. The functions $g_k(x)$ satisfy the inequality in (1.2); that is,

$$0 \leqslant g_k(x) \leqslant G_{k-1}(x) - G_{k-1}(x^{k-1}),$$

for $k = 2, 3, \ldots$. Consequently,

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

Assumption 4. There is a real number α with

$$\alpha \leqslant f(x),$$

for all x in R^J.

Assumption 5. Each x^k is in D.

Using these assumptions, we can conclude several things about the sequence $\{x^k\}$.

Proposition 6.1. *The sequence $\{f(x^k)\}$ is decreasing, and the sequence $\{g_k(x^k)\}$ converges to zero.*

Proof. We have

$$f(x^{k+1}) + g_{k+1}(x^{k+1}) = G_{k+1}(x^{k+1}) \leqslant G_{k+1}(x^k) = f(x^k) + g_{k+1}(x^k) = f(x^k).$$

Therefore,

$$f(x^k) - f(x^{k+1}) \geqslant g_{k+1}(x^{k+1}) \geqslant 0.$$

Since the sequence $\{f(x^k)\}$ is decreasing and bounded below by $f(\hat{x})$, the difference sequence must converge to zero. Therefore, the sequence $\{g_k(x^k)\}$ converges to zero. □

Theorem 6.1. *The sequence $\{f(x^k)\}$ converges to $f(\hat{x})$.*

Proof. Suppose that there is $\delta > 0$ with

$$f(x^k) \geqslant f(\hat{x}) + \delta,$$

for all k. Since $\hat{x}$ is in C, there is z in D with

$$f(x^k) \geqslant f(z) + \frac{\delta}{2},$$

for all k. From

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

we have

$$g_k(z) - g_{k+1}(z) \geqslant f(x^k) + g_k(x^k) - f(z) \geqslant f(x^k) - f(z) \geqslant \frac{\delta}{2} > 0.$$

This says that the nonnegative sequence $\{g_k(z)\}$ is decreasing, but that successive differences remain bounded away from zero, which cannot happen. □

Theorem 6.2. *Let the restriction of $f(x)$ to x in C have bounded level sets. Then the sequence $\{x^k\}$ is bounded, and $f(x^*) = f(\hat{x})$, for any cluster point x^*. If $\hat{x}$ is unique, $x^* = \hat{x}$ and $\{x^k\} \to \hat{x}$.*

Proof. From the previous theorem we have $f(x^*) = f(\hat{x})$, for all cluster points x^*. But, by uniqueness, $x^* = \hat{x}$, and so $\{x^k\} \to \hat{x}$. □

Corollary 6.1. *Let $f(x)$ be closed, proper and convex. If $\hat{x}$ is unique, the sequence $\{x^k\}$ converges to $\hat{x}$.*

Proof. Let $\iota_C(x)$ be the indicator function of the set C, that is, $\iota_C(x) = 0$, for all x in C, and $\iota_C(x) = +\infty$, otherwise. Then the function $g(x) = f(x) + \iota_C(x)$ is closed, proper and convex. If $\hat{x}$ is unique, then we have

$$\{x | f(x) + \iota_C(x) \leqslant f(\hat{x})\} = \{\hat{x}\}.$$

Therefore, one of the level sets of $g(x)$ is bounded and nonempty. It follows from Corollary 8.7.1 of [22] that every level set of $g(x)$ is bounded, so that the sequence $\{x^k\}$ is bounded. □

If $\hat{x}$ is not unique, we may still be able to prove convergence of the sequence $\{x^k\}$, for particular cases of SUMMA, as we shall see shortly.

7. Barrier-function methods (II)

We return now to the barrier-function methods, to show that they are particular cases of the SUMMA. The iterative step of the barrier-function method can be formulated as follows: minimize

$$f(x) + [(k-1)f(x) + b(x)] \tag{7.1}$$

to obtain x^k. Since, for $k = 2, 3, \ldots$, the function

$$(k-1)f(x) + b(x) \tag{7.2}$$

is minimized by x^{k-1}, the function

$$g_k(x) = (k-1)f(x) + b(x) - (k-1)f(x^{k-1}) - b(x^{k-1}) \tag{7.3}$$

is nonnegative, and x^k minimizes the function

$$G_k(x) = f(x) + g_k(x). \tag{7.4}$$

From

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

it follows that

$$G_k(x) - G_k(x^k) = kf(x) + b(x) - kf(x^k) - b(x^k) = g_{k+1}(x),$$

so that $g_{k+1}(x)$ satisfies the condition in (1.2). This shows that the barrier-function method is a particular case of SUMMA.

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

$$f(x) + \frac{1}{k}b(x) \tag{7.5}$$

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

From the results for SUMMA, we conclude that $\{f(x^k)\}$ is decreasing to $f(\hat{x})$, and that $\{g_k(x^k)\}$ converges to zero. From the nonnegativity of $g_k(x^k)$ we have that

$$(k-1)(f(x^k) - f(x^{k-1})) \geqslant b(x^{k-1}) - b(x^k).$$

Since the sequence $\{f(x^k)\}$ is decreasing, the sequence $\{b(x^k)\}$ must be increasing, but might not be bounded above.

If $\hat{x}$ is unique, and $f(x)$ has bounded level sets, then it follows, from our discussion of SUMMA, that $\{x^k\} \to \hat{x}$. Suppose now that $\hat{x}$ is not known to be unique, but can be chosen in D, so that $G_k(\hat{x})$ is finite for each k. From

$$f(\hat{x}) + \frac{1}{k}b(\hat{x}) \geqslant f(x^k) + \frac{1}{k}b(x^k)$$

we have

$$\frac{1}{k}(b(\hat{x}) - b(x^k)) \geqslant f(x^k) - f(\hat{x}) \geqslant 0,$$

so that

$$b(\hat{x}) - b(x^k) \geqslant 0,$$

for all k. If either f or b has bounded level sets, then the sequence $\{x^k\}$ is bounded and has a cluster point, x^* in C. It follows that $b(x^*) \leqslant b(\hat{x}) < +\infty$, so that x^* is in D. If we assume that $f(x)$ is convex and $b(x)$ is strictly convex on D, then we can show that x^* is unique in D, so that $x^* = \hat{x}$ and $\{x^k\} \to \hat{x}$.

To see this, assume, to the contrary, that there are two distinct cluster points x^* and x^{**} in D, with

$$\{x^{k_n}\} \to x^*,$$

and

$$\{x^{j_n}\} \to x^{**}.$$

Without loss of generality, we assume that

$$0 < k_n < j_n < k_{n+1},$$

for all n, so that

$$b(x^{k_n}) \leqslant b(x^{j_n}) \leqslant b(x^{k_{n+1}}).$$

Therefore,

$$b(x^*) = b(x^{**}) \leqslant b(\hat{x}).$$

From the strict convexity of $b(x)$ on the set D, and the convexity of $f(x)$, we conclude that, for $0 < \lambda < 1$ and $y = (1-\lambda)x^* + \lambda x^{**}$, we have $b(y) < b(x^*)$ and $f(y) \leqslant f(x^*)$. But, we must then have $f(y) = f(x^*)$. There must then be some k_n such that

$$G_{k_n}(y) = f(y) + \frac{1}{k_n}b(y) < f(x_{k_n}) + \frac{1}{k_n}b(x_{k_n}) = G_{k_n}(x^{k_n}).$$

But, this is a contradiction.

The following theorem summarizes what we have shown with regard to the barrier-function method.

Theorem 7.1. *Let $f : R^J \to (-\infty, +\infty]$ be a continuous function. Let $b(x) : R^J \to (0, +\infty]$ be a continuous function, with effective domain the nonempty set D. Let $\hat{x}$ minimize $f(x)$ over all x in $C = \overline{D}$. For each positive integer k, let x^k minimize the function $f(x) + \frac{1}{k}b(x)$. Then the sequence $\{f(x^k)\}$ is monotonically decreasing to the limit $f(\hat{x})$, and the sequence $\{b(x^k)\}$ is increasing. If $\hat{x}$ is unique, and $f(x)$ has bounded level sets, then the sequence $\{x^k\}$ converges to $\hat{x}$. In particular, if $\hat{x}$ can be chosen in D, if either $f(x)$ or $b(x)$ has bounded level sets, if $f(x)$ is convex and if $b(x)$ is strictly convex on D, then $\hat{x}$ is unique in D and $\{x^k\}$ converges to $\hat{x}$.*

Each step of the barrier method requires the minimization of the function $f(x) + \frac{1}{k}b(x)$. In practice, this must also be performed iteratively, with, say, the Newton–Raphson algorithm. It is important, therefore, that barrier functions be selected so that relatively few Newton–Raphson steps are needed to produce acceptable solutions to the main problem. For more on these issues see Renegar [21] and Nesterov and Nemirovski [20].

8. Penalty-function methods (II)

Let M be the non-empty closed set of all x for which the continuous function $p(x)$ attains its minimum value; this value need not be zero. Now we consider the problem of minimizing a continuous function $f(x) : R^J \rightarrow (-\infty, +\infty]$ over the closed set M. We assume that the constrained minimum of $f(x)$ is attained at some vector $\hat{x}$ in M. We also assume that the function $p(x)$ has bounded level sets, that is, for all $\gamma \geqslant 0$, the set $\{x|p(x) \leqslant \gamma\}$ is bounded.

For $k = 1, 2, \ldots$, let x^k be a minimizer of the function $f(x) + kp(x)$. As we shall see, we can formulate this penalty-function algorithm as a barrier-function iteration.

8.1. Penalty-function methods as barrier-function methods

In order to relate penalty-function methods to barrier-function methods, we note that minimizing $f(x) + kp(x)$ is equivalent to minimizing $p(x) + \frac{1}{k} f(x)$. This is the form of the barrier-function iteration, with $p(x)$ now in the role previously played by $f(x)$, and $f(x)$ now in the role previously played by $b(x)$. We are not concerned here with the effective domain of $f(x)$.

Now our assumption 2 simply says that there is a vector $\hat{x}$ at which $p(x)$ attains its minimum; so M is not empty. From our discussion of barrier-function methods, we know that the sequence $\{p(x^k)\}$ is decreasing to a limit $\hat{p} \geqslant p(\hat{x})$ and the sequence $\{f(x^k)\}$ is increasing. Since $p(x)$ has bounded level sets, the sequence $\{x^k\}$ is bounded; let x^* be an arbitrary cluster point. We then have $p(x^*) = \hat{p}$. It may seem odd that we are trying to minimize $f(x)$ over the set M using a sequence $\{x^k\}$ with $\{f(x^k)\}$ increasing, but remember that these x^k are not in M.

We now show that $f(x^*) = f(\hat{x})$; this does not follow from our previous discussion of barrier-function methods.

Let $s(x) = p(x) - p(\hat{x})$, so that $s(x) \geqslant 0$ and $s(\hat{x}) = 0$. For each k, let

$$T_k(x) = f(x) + ks(x) = f(x) + kp(x) - kp(\hat{x}).$$

Then x^k minimizes $T_k(x)$.

Lemma 8.1. *The sequence $\{T_k(x^k)\}$ is increasing to some limit $\gamma \leqslant f(\hat{x})$.*

Proof. Because the penalty function $s(x)$ is nonnegative, we have

$$T_k(x^k) \leqslant T_k(x^{k+1}) \leqslant T_k(x^{k+1}) + s(x^{k+1}) = T_{k+1}(x^{k+1}).$$

We also have

$$f(\hat{x}) = f(\hat{x}) + ks(\hat{x}) = T_k(\hat{x}) \geqslant T_k(x^k),$$

for all k. □

Lemma 8.2. *For all cluster points x^* of $\{x^k\}$ we have $s(x^*) = 0$, so that $p(x^*) = p(\hat{x})$ and x^* is in M.*

Proof. For each k we have

$$\alpha + ks(x^k) \leqslant f(x^k) + ks(x^k) = T_k(x^k) \leqslant f(\hat{x}),$$

so that

$$0 \leqslant ks(x^k) \leqslant f(\hat{x}) - \alpha,$$

for all k. It follows that $\{s(x^k)\}$ converges to zero. By the continuity of $s(x)$, we conclude that $s(x^*) = 0$, so x^* is in M. □

Lemma 8.3. *For all cluster points x^* of the sequence $\{x^k\}$ we have $f(x^*) = f(\hat{x})$, so x^* minimizes $f(x)$ over x in M.*

Proof. Let $\{x^{k_n}\} \to x^*$. We have

$$f(x^*) = f(x^*) + s(x^*) = \lim_{n\to+\infty} (f(x^{k_n}) + s(x^{k_n}))$$
$$\leqslant \lim_{n\to+\infty} (f(x^{k_n}) + k_n s(x^{k_n})) \leqslant f(\hat{x}).$$

Since x^* is in M, it follows that $f(x^*) = f(\hat{x})$. □

To assert that the sequence $\{x^k\}$ itself converges, we would need to make additional assumptions. For example, if the minimizer of $f(x)$ over x in M is unique, then the sequence $\{x^k\}$ has $\hat{x}$ for its only cluster point, so must converge to $\hat{x}$.

The following theorem summarizes what we have shown with regard to penalty-function methods.

Theorem 8.1. *Let $f : R^J \to (-\infty, +\infty]$ be a continuous function. Let $p(x) : R^J \to R$ be a continuous function, with bounded level sets, and M be the set of all $\tilde{x}$ such that $p(\tilde{x}) \leqslant p(x)$ for all x in R^J. Let $\hat{x}$ in M minimize $f(\tilde{x})$ over all $\tilde{x}$ in M. For each positive integer k, let x^k minimize the function $f(x) + kp(x)$. Then the sequence $\{f(x^k)\}$ is monotonically increasing to the limit $f(\hat{x})$, and the sequence $\{p(x^k)\}$ is decreasing to $p(\hat{x})$. If $\hat{x}$ is unique, which happens, for example, if $f(x)$ is strictly convex on M, then the sequence $\{x^k\}$ converges to $\hat{x}$.*

9. The proximal minimization algorithm (II)

We show now that assumption 3 holds, so that the PMA is a particular case of the SUMMA. We remind the reader that $f(x)$ is now assumed to be convex and differentiable, so that the Bregman distance $D_f(x, z)$ is defined and nonnegative, for all x in D and z in int D.

Lemma 9.1. *For each k we have*

$$G_k(x) = G_k(x^k) + D_f(x, x^k) + D_h(x, x^k). \tag{9.1}$$

Proof. Since x^k minimizes $G_k(x)$ within the set D, we have

$$0 = \nabla f(x^k) + \nabla h(x^k) - \nabla h(x^{k-1}). \tag{9.2}$$

Then

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

Now substitute, using equation (9.2), and use the definition of Bregman distances. □

It follows from lemma 9.1 that

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

so assumption 3 holds.

From the discussion of the SUMMA we know that $\{f(x^k)\}$ is monotonically decreasing to $f(\hat{x})$. As we noted previously, if the sequence $\{x^k\}$ is bounded, and $\hat{x}$ is unique, we can conclude that $\{x^k\} \to \hat{x}$.

Suppose that $\hat{x}$ is not known to be unique, but can be chosen in D; this will be the case, of course, whenever D is closed. Then $G_k(\hat{x})$ is finite for each k. From the definition of $G_k(x)$ we have

$$G_k(\hat{x}) = f(\hat{x}) + D_h(\hat{x}, x^{k-1}). \tag{9.3}$$

From equation (9.1) we have

$$G_k(\hat{x}) = G_k(x^k) + D_f(\hat{x}, x^k) + D_h(\hat{x}, x^k), \tag{9.4}$$

so that

$$G_k(\hat{x}) = f(x^k) + D_h(x^k, x^{k-1}) + D_f(\hat{x}, x^k) + D_h(\hat{x}, x^k). \tag{9.5}$$

Therefore,

$$D_h(\hat{x}, x^{k-1}) - D_h(\hat{x}, x^k) = f(x^k) - f(\hat{x}) + D_h(x^k, x^{k-1}) + D_f(\hat{x}, x^k). \tag{9.6}$$

It follows that the sequence $\{D_h(\hat{x}, x^k)\}$ is decreasing and that the sequence $\{D_f(\hat{x}, x^k)\}$ converges to 0. If either the function $f(x)$ or the function $D_h(\hat{x}, \cdot)$ has bounded level sets, then the sequence $\{x^k\}$ is bounded, has cluster points x^* in C, and $f(x^*) = f(\hat{x})$, for every x^*. We now show that $\hat{x}$ in D implies that x^* is also in D, whenever h is a Bregman–Legendre function.

Let x^* be an arbitrary cluster point, with $\{x^{k_n}\} \to x^*$. If $\hat{x}$ is not in int D, then, by property B2 of Bregman–Legendre functions, we know that

$$D_h(x^*, x^{k_n}) \to 0,$$

so x^* is in D. Then the sequence $\{D_h(x^*, x^k)\}$ is decreasing. Since a subsequence converges to zero, we have $\{D_h(x^*, x^k)\} \to 0$. From property R5, we conclude that $\{x^k\} \to x^*$.

If $\hat{x}$ is in int D, but x^* is not, then $\{D_h(\hat{x}, x^k)\} \to +\infty$, by property R2. But, this is a contradiction; therefore x^* is in D. Once again, we conclude that $\{x^k\} \to x^*$.

Now we summarize our results for the PMA. Let $f : R^J \to (-\infty, +\infty]$ be closed, proper, convex and differentiable. Let h be a closed proper convex function, with effective domain D, that is differentiable on the nonempty open convex set int D. Assume that $f(x)$ is finite on $C = \overline{D}$ and attains its minimum value on C at $\hat{x}$. For each positive integer k, let x^k minimize the function $f(x) + D_h(x, x^{k-1})$. Assume that each x^k is in the interior of D.

Theorem 9.1. *If the restriction of $f(x)$ to x in C has bounded level sets and $\hat{x}$ is unique, then the sequence $\{x^k\}$ converges to $\hat{x}$.*

Theorem 9.2. *If $h(x)$ is a Bregman–Legendre function and $\hat{x}$ can be chosen in D, then $\{x^k\} \to x^*$, x^* in D, with $f(x^*) = f(\hat{x})$.*

9.1. The method of Auslander and Teboulle

The method of Auslander and Teboulle described in a previous section seems not to be a particular case of SUMMA. However, we can adapt the proof of theorem 6.1 to prove the analogous result for their method. Once again, we assume that $f(\hat{x}) \leqslant f(x)$, for all x in C.

Theorem 9.3. *For $k = 2, 3, \ldots$, let x^k minimize the function*

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

If the distance d has an induced proximal distance H, then $\{f(x^k)\} \to f(\hat{x})$.

Proof. First, we show that the sequence $\{f(x^k)\}$ is decreasing. We have

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

from which we conclude that the sequence $\{f(x^k)\}$ is decreasing and the sequence $\{d(x^k, x^{k-1})\}$ converges to zero.

Now suppose that

$$f(x^k) \geqslant f(\hat{x}) + \delta,$$

for some $\delta > 0$ and all k. Since $\hat{x}$ is in C, there is z in D with

$$f(x^k) \geqslant f(z) + \frac{\delta}{2},$$

for all k. Since x^k minimizes $F_k(x)$, it follows that

$$0 = \nabla f(x^k) + \nabla_1 d(x^k, x^{k-1}).$$

Using the convexity of the function $f(x)$ and the fact that H is an induced proximal distance, we have

$$\begin{aligned} 0 < \frac{\delta}{2} &\leqslant f(x^k) - f(z) \leqslant \langle -\nabla f(x^k), z - x^k \rangle \\ &= \langle \nabla_1 d(x^k, x^{k-1}), z - x^k \rangle \leqslant H(z, x^{k-1}) - H(z, x^k). \end{aligned}$$

Therefore, the nonnegative sequence $\{H(z, x^k)\}$ is decreasing, but its successive differences remain bounded below by $\frac{\delta}{2}$, which is a contradiction. □

It is interesting to note that the Auslander–Teboulle approach places a restriction on the function $d(x, y)$, the existence of the induced proximal distance H, that is unrelated to the objective function $f(x)$, but this condition is helpful only for convex $f(x)$. In contrast, the SUMMA approach requires that

$$0 \leqslant g_{k+1}(x) \leqslant G_k(x) - G_k(x^k),$$

which involves the $f(x)$ being minimized, but does not require that this $f(x)$ be convex.

9.2. Bregman-dominated distances

Assume that, for each fixed a, the function $g(x) = d(x, a)$ is such that the associated Bregman distance $D_a(c, b)$ can be defined. Then

$$D_a(c, b) = g(c) - g(b) - \langle \nabla g(b), c - b \rangle \geqslant 0, \tag{9.7}$$

for all b and c. Therefore,

$$D_a(c, b) = d(c, a) - d(b, a) - \langle \nabla_1 d(b, a), c - b \rangle \geqslant 0, \tag{9.8}$$

for all b and c. Say that the distance d is *Bregman dominated* if

$$D_a(c, b) \geqslant d(c, b), \tag{9.9}$$

for all a, b and c. It follows then that

$$d(c, a) - d(b, a) - \langle \nabla_1 d(b, a), c - b \rangle \geqslant d(c, b), \tag{9.10}$$

or

$$\langle \nabla_1 d(b, a), c - b \rangle \leqslant d(c, a) - d(c, b) - d(b, a) \leqslant d(c, a) - d(c, b). \tag{9.11}$$

Consequently, the choice of $H = d$ satisfies the inequality in (4.5).

For each k, let $D_k(x, y) = D_a(x, y)$, for the choice $a = x^{k-1}$. Since x^{k-1} minimizes the function $d(x, x^{k-1})$, we have

$$\nabla_1 d(x^{k-1}, x^{k-1}) = 0,$$

and so

$$D_k(x, x^{k-1}) = d(x, x^{k-1}).$$

Therefore, x^k minimizes the function

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

From lemma 9.1, we conclude that

$$G_k(x) - G_k(x^k) = D_f(x, x^k) + D_k(x, x^k) \geqslant D_k(x, x^k), \tag{9.12}$$

assuming, of course, that f is convex.

If the distance d is Bregman dominated, then we have

$$G_k(x) - G_k(x^k) \geqslant D_k(x, x^k) \geqslant d(x, x^k), \tag{9.13}$$

so the iteration is a particular case of SUMMA.

10. The simultaneous MART (II)

It follows from the identities established in [6] that the SMART can also be formulated as a particular case of the SUMMA.

10.1. The SMART as a case of SUMMA

We show now that the SMART is a particular case of the SUMMA. The following lemma is helpful in that regard.

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

$$\mathrm{KL}(x, z) = \mathrm{KL}(x_+, z_+) + \mathrm{KL}\left(x, \frac{x_+}{z_+} z\right). \tag{10.1}$$

From the identities established for the SMART in [6], we know that the iterative step of SMART can be expressed as follows: minimize the function

$$G_k(x) = \mathrm{KL}(Px, y) + \mathrm{KL}(x, x^{k-1}) - \mathrm{KL}(Px, Px^{k-1}) \tag{10.2}$$

to obtain x^k. According to lemma 10.1, the quantity

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

is nonnegative, since $s_j = 1$. The $g_k(x)$ are defined for all nonnegative x; that is, the set D is the closed nonnegative orthant in R^J. Each x^k is a positive vector.

It was shown in [6] that

$$G_k(x) = G_k(x^k) + \mathrm{KL}(x, x^k), \tag{10.3}$$

from which it follows immediately that assumption 3 holds for the SMART.

Because the SMART is a particular case of the SUMMA, we know that the sequence $\{f(x^k)\}$ is monotonically decreasing to $f(\hat{x})$. It was shown in [6] that if $y = Px$ has no nonnegative solution and the matrix P and every submatrix obtained from P by removing columns has full rank, then $\hat{x}$ is unique; in that case, the sequence $\{x^k\}$ converges to $\hat{x}$. As we shall see, the SMART sequence always converges to a nonnegative minimizer of $f(x)$. To establish this, we reformulate the SMART as a particular case of the PMA.

10.2. The SMART as a case of the PMA

We take $F(x)$ to be the function

$$F(x) = \sum_{j=1}^J x_j \log x_j. \tag{10.4}$$

Then

$$D_F(x, z) = \mathrm{KL}(x, z). \tag{10.5}$$

For nonnegative x and z in $\mathcal{X}$, we have

$$D_f(x, z) = \mathrm{KL}(Px, Pz). \tag{10.6}$$

Lemma 10.2. $D_F(x, z) \geqslant D_f(x, z)$.

Proof. We have

$$\begin{aligned} D_F(x, z) &\geqslant \sum_{j=1}^{J} \mathrm{KL}(x_j, z_j) \geqslant \sum_{j=1}^{J} \sum_{i=1}^{I} \mathrm{KL}(P_{ij}x_j, P_{ij}z_j) \\ &\geqslant \sum_{i=1}^{I} KL((Px)_i, (Pz)_i) = \mathrm{KL}(Px, Pz). \end{aligned} \tag{10.7}$$

□

Then we let $h(x) = F(x) - f(x)$; then $D_h(x, z) \geqslant 0$ for nonnegative x and z in $\mathcal{X}$. The iterative step of the SMART is to minimize the function

$$f(x) + D_h(x, x^{k-1}). \tag{10.8}$$

So the SMART is a particular case of the PMA.

The function $h(x) = F(x) - f(x)$ is finite on D the nonnegative orthant of R^J, and differentiable on the interior, so $C = D$ is closed in this example. Consequently, $\hat{x}$ is necessarily in D. From our earlier discussion of the PMA, we can conclude that the sequence $\{D_h(\hat{x}, x^k)\}$ is decreasing and the sequence $\{D_f(\hat{x}, x^k)\} \to 0$. Since the function $\mathrm{KL}(\hat{x}, \cdot)$ has bounded level sets, the sequence $\{x^k\}$ is bounded, and $f(x^*) = f(\hat{x})$, for every cluster point. Therefore, the sequence $\{D_h(x^*, x^k)\}$ is decreasing. Since a subsequence converges to zero, the entire sequence converges to zero. The convergence of $\{x^k\}$ to x^* follows from basic properties of the KL distance.

From the fact that $\{D_f(\hat{x}, x^k)\} \to 0$, we conclude that $P\hat{x} = Px^*$. Equation (9.6) now tells us that the difference $D_h(\hat{x}, x^{k-1}) - D_h(\hat{x}, x^k)$ depends only on $P\hat{x}$, and not directly on $\hat{x}$. Therefore, the difference $D_h(\hat{x}, x^0) - D_h(\hat{x}, x^*)$ also depends only on $P\hat{x}$ and not directly on $\hat{x}$. Minimizing $D_h(\hat{x}, x^0)$ over nonnegative minimizers $\hat{x}$ of $f(x)$ is therefore equivalent to minimizing $D_h(\hat{x}, x^*)$ over the same vectors. But the solution to the latter problem is obviously $\hat{x} = x^*$. Thus we have shown that the limit of the SMART is the nonnegative minimizer of $\mathrm{KL}(Px, y)$ for which the distance $\mathrm{KL}(x, x^0)$ is minimized.

The following theorem summarizes the situation with regard to the SMART.

Theorem 10.1. *In the consistent case the SMART converges to the unique nonnegative solution of* $y = Px$ *for which the distance* $\mathrm{KL}(x, x^0)$ *is minimized. In the inconsistent case it converges to the unique nonnegative minimizer of the distance* $\mathrm{KL}(Px, y)$ *for which* $\mathrm{KL}(x, x^0)$ *is minimized; if* P *and every matrix derived from* P *by deleting columns has full rank then there is a unique nonnegative minimizer of* $\mathrm{KL}(Px, y)$ *and at most* $I - 1$ *of its entries are nonzero.*

10.3. The EMML algorithm

The expectation maximization maximum likelihood (EMML) algorithm minimizes the function $f(x) = \mathrm{KL}(y, Px)$ over x in $\mathcal{X}$. In [12] the EMML algorithm and the SMART are developed in tandem to reveal how closely related these two methods are. There, the EMML algorithm is derived using alternating minimization, in which the vector x^k is the one

for which the function $\mathrm{KL}(r(x^{k-1}), q(x))$ is minimized. When we try to put the EMML into the framework of SUMMA, we find that x^k minimizes the function

$$G_k(x) = f(x) + \mathrm{KL}(r(x^{k-1}), r(x)), \tag{10.9}$$

over all positive vectors x. However, the functions

$$g_k(x) = \mathrm{KL}(r(x^{k-1}), r(x)) \tag{10.10}$$

appear not to satisfy the condition in (1.2). It appears not to be true that the EMML is a particular case of SUMMA, even though it is true that $\{f(x^k)\}$ does converge monotonically to $f(\hat{x})$ and $\{x^k\}$ does converge to a nonnegative minimizer of $f(x)$. The obvious conjecture is that the EMML is an example of a wider class of sequential unconstrained minimization algorithms for which a nice theory of convergence still holds.

In the next section we present a variant of the SMART, designed to incorporate bounds on the entries of the vector x.

11. Minimizing $\mathrm{KL}(Px, y)$ with bounds on x

Let $a_j < b_j$, for each j. Let $\mathcal{X}_{ab}$ be the set of all vectors x such that $a_j \leqslant x_j \leqslant b_j$, for each j. Now, we seek to minimize $f(x) = \mathrm{KL}(Px, y)$, over all vectors x in $\mathcal{X} \cap \mathcal{X}_{ab}$. We let

$$F(x) = \sum_{j=1}^{J}((x_j - a_j)\log(x_j - a_j) + (b_j - x_j)\log(b_j - x_j)). \tag{11.1}$$

Then we have

$$D_F(x, z) = \sum_{j=1}^{J}(\mathrm{KL}(x_j - a_j, z_j - a_j) + \mathrm{KL}(b_j - x_j, b_j - z_j)), \tag{11.2}$$

and, as before,

$$D_f(x, z) = \mathrm{KL}(Px, Pz). \tag{11.3}$$

Lemma 11.1. *For any $c > 0$, with $a \geqslant c$ and $b \geqslant c$, we have* $\mathrm{KL}(a - c, b - c) \geqslant \mathrm{KL}(a, b)$.

Proof. Let $g(c) = \mathrm{KL}(a - c, b - c)$ and differentiate with respect to c, to obtain

$$g'(c) = \frac{a - c}{b - c} - 1 - \log\left(\frac{a - c}{b - c}\right) \geqslant 0. \tag{11.4}$$

We see then that the function $g(c)$ is increasing with c. □

As a corollary of lemma 11.1, we have

Lemma 11.2. *Let $a = (a_1, \ldots, a_J)^T$, and x and z in $\mathcal{X}$ with $(Px)_i \geqslant (Pa)_i$, $(Pz)_i \geqslant (Pa)_i$, for each i. Then* $\mathrm{KL}(Px, Pz) \leqslant \mathrm{KL}(Px - Pa, Pz - Pa)$.

Lemma 11.3. $D_F(x, z) \geqslant D_f(x, z)$.

Proof. We can easily show that $D_F(x, z) \geqslant \mathrm{KL}(Px - Pa, Pz - Pa) + \mathrm{KL}(Pb - Px, Pb - Pz)$, along the lines used previously. Then, from lemma 11.2, we have $\mathrm{KL}(Px - Pa, Pz - Pa) \geqslant \mathrm{KL}(Px, Pz) = D_f(x, z)$. □

Once again, we let $h(x) = F(x) - f(x)$, which is finite on the closed convex set $\mathcal{X} \cap \mathcal{X}_{ab}$. At the kth step of this algorithm we minimize the function

$$f(x) + D_h(x, x^{k-1}) \tag{11.5}$$

to obtain x^k.

Solving for x_j^k, we obtain

$$x_j^{k+1} = \alpha_j^k a_j + \left(1 - \alpha_j^k\right) b_j, \tag{11.6}$$

where

$$\left(\alpha_j^k\right)^{-1} = 1 + \left(\frac{x_j^{k-1} - a_j}{b_j - x_j^{k-1}}\right) \exp\left(\sum_{i=1}^{I} P_{ij} \log(y_i/(Px^{k-1})_i)\right). \tag{11.7}$$

Since the restriction of $f(x)$ to $\mathcal{X} \cap \mathcal{X}_{ab}$ has bounded level sets, the sequence $\{x^k\}$ is bounded and has cluster points. If $\hat{x}$ is unique, then $\{x^k\} \to \hat{x}$.

This algorithm is closely related to those presented in [7].

12. Related Work

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 [19] is one example of this.

12.1. The Moreau envelope

Proposition 12.1. *The infimum of $m_f(z)$, over all z, is the same as the infimum of $f(x)$, over all x.*

Proof. We have

$$\begin{aligned}\inf_z m_f(z) &= \inf_z \inf_x \left\{f(x) + \tfrac{1}{2}||x - z||_2^2\right\} \\ &= \inf_x \inf_z \left\{f(x) + \tfrac{1}{2}||x - z||_2^2\right\} = \inf_x \left\{f(x) + \tfrac{1}{2} \inf_z ||x - z||_2^2\right\} = \inf_x f(x).\end{aligned}$$ □

Later, we shall show that the minimizers of $m_f(z)$ and $f(x)$ are the same, as well.

Both m_f and m_{f^*} are convex and differentiable. 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, $f(x) = \psi_C(x)$ 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.

The operators $\text{prox}_f : z \to \text{prox}_f z$ are *proximal operators*. These operators generalize the projections onto convex sets, and, like those operators, are firmly non-expansive [14].

The support function of the convex set C is $\sigma_C(x) = \sup_{u \in C} \langle x, u \rangle$. It is easy to see that $\sigma_C = \psi_C^*$. For $f^*(z) = \sigma_C(z)$, we can find $m_{f^*} z$ using Moreau's theorem ([22], p 338).

12.2. Moreau's theorem and applications

Moreau's theorem generalizes the decomposition of members of R^J with respect to a subspace. For a proof, see the book by Rockafellar [22].

Theorem 12.1 (Moreau's theorem)**.** *Let f be a closed, proper, convex function. Then*

$$m_f z + m_{f^*} z = \tfrac{1}{2}||z||^2, \tag{12.1}$$

and

$$\text{prox}_f z + \text{prox}_{f^*} z = z. \tag{12.2}$$

In addition, we have

$$\begin{aligned} &\text{prox}_{f^*} z \in \partial f(\text{prox}_f z), \\ &\text{prox}_{f^*} z = \nabla m_f(z), \qquad \text{and} \\ &\text{prox}_f z = \nabla m_{f^*}(z). \end{aligned} \tag{12.3}$$

Since $\sigma_C = \psi_C^*$, we have

$$\text{prox}_{\sigma_C} z = z - \text{prox}_{\psi_C} z = z - P_C z. \tag{12.4}$$

The following proposition illustrates the usefulness of these concepts.

Proposition 12.2. *The minimizers of m_f and the minimizers of f are the same.*

Proof. From Moreau's theorem we know that

$$\nabla m_f(z) = \text{prox}_{f^*} z = z - \text{prox}_f z, \tag{12.5}$$

so $\nabla m_f z = 0$ is equivalent to $z = \text{prox}_f z$. □

12.3. Iterative minimization of $m_f z$

Because the minimizers of m_f are also minimizers of f, we can find global minimizers of f using standard iterative methods, such as gradient descent, on m_f. The gradient descent iterative step has the form

$$x^{k+1} = x^k - \gamma_k \nabla m_f(x^k). \tag{12.6}$$

We know from Moreau's theorem that

$$\nabla m_f z = \text{prox}_{f^*} z = z - \text{prox}_f z, \tag{12.7}$$

so that equation (12.6) can be written as

$$\begin{aligned} x^{k+1} &= x^k - \gamma_k (x^k - \text{prox}_f x^k) \\ &= (1 - \gamma_k) x^k + \gamma_k \,\text{prox}_f x^k. \end{aligned} \tag{12.8}$$

Because

$$x^k - \text{prox}_f x^k \in \partial f(\text{prox}_f x^k), \tag{12.9}$$

the iteration in equation (12.8) has the increment

$$x^{k+1} - x^k \in -\gamma_k \partial f(x^{k+1}), \tag{12.10}$$

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

$$x^{k+1} - x^k = -\gamma_k \nabla f(x^k). \tag{12.11}$$

It follows from the definition of $\partial f(x^{k+1})$ that $f(x^k) \geqslant f(x^{k+1})$ for the iteration in equation (12.8).

12.4. Forward-backward splitting

In [14] the authors 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

$$0 \in \partial f(x) = \partial f_1(x) + \nabla f_2(x), \tag{12.12}$$

so we have

$$-\gamma \nabla f_2(x) \in \gamma \partial f_1(x), \tag{12.13}$$

for any $\gamma > 0$. Therefore

$$x - \gamma \nabla f_2(x) - x \in \gamma \partial f_1(x). \tag{12.14}$$

From equation (12.14) we conclude that

$$x = \text{prox}_{\gamma f_1}(x - \gamma \nabla f_2(x)). \tag{12.15}$$

This suggests an algorithm with the iterative step

$$x^{k+1} = \text{prox}_{\gamma f_1}(x^k - \gamma \nabla f_2(x^k)). \tag{12.16}$$

In order to guarantee convergence, γ is chosen to lie in the interval $(0, 2/\lambda)$. It is also possible to allow γ to vary with the k. This is called the *forward–backward splitting* approach. As noted in [14], the forward–backward splitting approach has, as a particular case, the CQ algorithm of [9, 10].

12.5. Generalizing the Moreau envelope

The Moreau envelope involves the infimum of the function

$$f(x) + \tfrac{1}{2}||x - z||_2^2. \tag{12.17}$$

Consequently, the Moreau envelope can be generalized in various ways, either by changing the $\frac{1}{2}$ to a variable parameter, or replacing the Euclidean distance by a more general *distance measure*.

For real $\lambda > 0$, the Moreau–Yosida approximation of index λ [1] is the function

$$F_\lambda(z) = \inf_x \left\{ f(x) + \frac{1}{2\lambda}||x - z||_2^2 \right\}. \tag{12.18}$$

For fixed λ, the theory is much the same as for the Moreau envelope [1, 2]. For fixed λ, $F_\lambda(z)$ can be viewed as an approximate minimization of $f(x)$, involving regularization based on an additive penalty term. If $z = 0$, then $F_\lambda(0)$ is a norm-constrained minimization of $f(x)$.

12.6. Proximity operators using Bregman distances

Several authors have extended Moreau's results by replacing the Euclidean squared distance with a Bregman distance. Let h be a closed proper convex function that is differentiable on the nonempty set int D. The corresponding *Bregman distance* $D_h(x, z)$ is defined for $x \in R^J$ and $z \in$ int D by

$$D_h(x, z) = h(x) - h(z) - \langle \nabla h(z), x - z \rangle. \tag{12.19}$$

Note that $D_h(x, z) \geqslant 0$ always and that $D_h(x, z) = +\infty$ is possible. If h is essentially strictly convex then $D_h(x, z) = 0$ implies that $x = z$.

12.6.1. Teboulle's entropic proximal mappings. Teboulle [24] considers the function

$$R(x, z) = f(x) + \epsilon D_h(x, z), \tag{12.20}$$

and shows that, with certain restrictions on f and h, the function $R(\cdot, z)$ attains its minimum value, $R_\epsilon(z)$, at a unique $x = E_h(f, z)$. He then generalizes Moreau's theorem, proving that the operator $E_h(f, \cdot)$ has properties analogous to the proximity operators $\text{prox}_f(\cdot)$. He then demonstrates that several nonlinear programming problems can be formulated using such functions $R(x, z)$. He is primarily concerned with the behavior of $R_\epsilon(z)$, as z varies, and not as ϵ varies.

Teboulle's method relies on Fenchel's Duality theorem [22], and therefore requires the conjugate of the function $g(x) = D_h(x, z)$. As he shows,

$$g^*(y) = h^*(y + \nabla h(z)) - h^*(\nabla h(z)). \tag{12.21}$$

His main result requires the joint convexity of the function $D_h(x, z)$.

12.6.2. The proximal minimization of Censor and Zenios. Censor and Zenios [13] also consider $R(x, z)$. They are less interested in the properties of the operator $E_h(f, \cdot)$ and more interested in the behavior of their PMD iterative algorithm defined by

$$x^{k+1} = \text{argmin}(f(x) + D_h(x, x^k)). \tag{12.22}$$

In their work, the function h is a Bregman function with zone S. They show that, subject to certain assumptions, if the function f has a minimizer within the closure of S, then the PMD iterates converge to such a minimizer. It is true that their method and results are somewhat more general, in that they consider also the minimizers of $R(x, z)$ over another closed convex set X; however, this set X is unrelated to the function h.

The PMA presented here has the same iterative step as the PMD method of Censor and Zenios. However, the assumptions about f and h are different, and our theorem asserts convergence of the iterates to a constrained minimizer of f over $C = \overline{D}$, whenever such a minimizer exists. In other words, we solve a constrained minimization problem, whereas Censor and Zenios solve the unconstrained minimization problem, under a restrictive assumption on the location of minimizers of f, and more restrictive assumptions on $h(x)$.

12.6.3. Alternating Bregman proximity operators. Recent work by Bauschke, Combettes and Noll [5] extends the proximal minimization idea to include the use of the Bregman distance with the variable vector in both the left and right positions, that is, minimizing functions of x involving first $D_f(x, x^{k-1})$ and then $D_f(x^k, x)$. This approach leads to alternating minimization algorithms for minimizing objective functions of the form

$$\varphi(x) + \psi(y) + D_f(x, y).$$

13. Computation

As we noted previously, we do not address computational issues in any detail in this paper. Nevertheless, it cannot be ignored that both equation (5.2) for the SMART and equations (11.6) and (11.7) for the generalized SMART provide easily calculated iterates, in contrast to other examples of SUMMA. At the same time, showing that these two algorithms are particular cases of SUMMA requires the introduction of functions $G_k(x)$ that appear to be quite ad hoc. The purpose of this section is to motivate these choices of $G_k(x)$ and to indicate how other analogous computationally tractable SUMMA iterative schemes may be derived.

13.1. Landweber's algorithm

Suppose that A is a real $I \times J$ matrix and we wish to obtain a least-squares solution $\hat{x}$ of $Ax = b$ by minimizing the function

$$f(x) = \tfrac{1}{2}||Ax - b||^2.$$

We know that

$$(A^T A)\hat{x} = A^T b, \tag{13.1}$$

so, in a sense, the problem is solved. However, in many applications, the dimensions I and J are quite large, perhaps in the tens of thousands, as in some image reconstruction problems. Solving equation (13.1), and even calculating $A^T A$, can be prohibitively expensive. In such cases, we turn to iterative methods, not necessarily to incorporate constraints on x, but to facilitate calculation. Landweber's algorithm is one such iterative method for calculating a least-squares solution.

The iterative step of Landweber's algorithm is

$$x^k = x^{k-1} - \gamma A^T(Ax^{k-1} - b). \tag{13.2}$$

The sequence $\{x^k\}$ converges to the least-squares solution closest to x^0, for any choice of γ in the interval $(0, 2/\rho(A^T A))$, where $\rho(A^T A)$, the spectral radius of $A^T A$, is its largest eigenvalue; this is a consequence of the Krasnoselskii–Mann theorem (see, for example, [10]).

It is easy to verify that the x^k given by equation (13.2) is the minimizer of the function

$$G_k(x) = \frac{1}{2}||Ax - b||^2 + \frac{1}{2\gamma}||x - x^{k-1}||^2 - \frac{1}{2}||Ax - Ax^{k-1}||^2, \tag{13.3}$$

that, for γ in the interval $(0, 1/\rho(A^T A))$, the iteration in equation (13.2) is a particular case of SUMMA, and

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

The similarity between the $G_k(x)$ in equation (13.3) and that in equation (10.2) is not accidental and both are particular cases of a more general iterative scheme involving proximal minimization.

13.2. Extending the PMA

The proximal minimization algorithm (PMA) requires us to minimize the function $G_k(x)$ given by equation (4.2) to obtain x^k. How x^k may be calculated was not addressed previously. Suppose, instead of minimizing $G_k(x)$ in equation (4.2), we minimize

$$G_k(x) = f(x) + D_h(x, x^{k-1}) - D_f(x, x^{k-1}), \tag{13.4}$$

with the understanding that $f(x)$ is convex and

$$D_h(x, z) - D_f(x, z) \geqslant 0,$$

for all appropriate x and z. The next iterate x^k satisfies the equation

$$0 = \nabla h(x^k) - \nabla h(x^{k-1}) + \nabla f(x^{k-1}), \tag{13.5}$$

so that

$$\nabla h(x^k) = \nabla h(x^{k-1}) - \nabla f(x^{k-1}). \tag{13.6}$$

This iterative scheme is the *interior-point algorithm* (IPA) presented in [8]. If the function $h(x)$ is chosen carefully, then we can solve for x^k easily. The Landweber algorithm, the SMART, and the generalized SMART are all particular cases of this IPA.

Using lemma 9.1, we can show that

$$G_k(x) - G_k(x^k) = D_h(x, x^k), \tag{13.7}$$

for all appropriate x, so that the IPA is a particular case of SUMMA. We consider now several other examples.

If we let $h(x) = \frac{1}{2\gamma}||x||^2$ in equation (13.4), the iteration becomes

$$x^k = x^{k-1} - \gamma \nabla f(x^{k-1}). \tag{13.8}$$

If, for example, the operator ∇f is L-Lipschitz continuous, that is,

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

then, for any γ in the interval $(0, 1/2L)$, we have

$$\frac{1}{2\gamma}||x - z||^2 \geqslant L||x - z||^2 \geqslant \langle \nabla f(x) - \nabla f(z), x - z\rangle = D_f(x, z) + D_f(z, x) \geqslant D_f(x, z).$$

Therefore, this iteration is a particular case of SUMMA. It should be noted that, in this case, the Krasnoselskii–Mann theorem gives convergence for any γ in the interval $(0, 2/L)$.

Finally, we consider what happens if we replace the Euclidean norm with that induced by the local geometry derived from f itself. More specifically, let us take

$$h(x) = \tfrac{1}{2}x^T \nabla^2 f(x^{k-1})x,$$

so that

$$D_h(x, x^{k-1}) = \tfrac{1}{2}(x - x^{k-1})^T \nabla^2 f(x^{k-1})(x - x^{k-1}).$$

Then the IPA iterate x^k becomes

$$x^k = x^{k-1} - \nabla^2 f(x^{k-1})^{-1} \nabla f(x^{k-1}), \tag{13.9}$$

which is the Newton–Raphson iteration. Using the SUMMA framework to study the Newton–Raphson method is work in progress.

Algorithms such as Landweber's and SMART can be slow to converge. It is known that convergence can often be accelerated using incremental gradient (partial gradient, block-iterative, ordered-subset) methods. Using the SUMMA framework to study such incremental gradient methods as the algebraic reconstruction technique (ART), its multiplicative version (MART), and other block-iterative methods is also the subject of on-going work.

Appendix: Bregman–Legendre functions

In [4] Bauschke and Borwein show convincingly that the Bregman–Legendre functions provide the proper context for the discussion of Bregman distances and Bregman projections onto closed convex sets. The summary here follows closely the discussion given in [4].

A.1. Basic notions

A convex function $h : R^J \to [-\infty, +\infty]$ is *proper* if there is no x with $h(x) = -\infty$ and some x with $h(x) < +\infty$. The *effective domain* of h is dom $h = \{x|h(x) < +\infty\}$. A proper convex function h is *closed* if it is lower semi-continuous, that is, if $h(x) = \liminf h(y)$, as $y \to x$. A convex function is continuous on any relatively open convex subset of dom h.

The *subdifferential* of h at x is the set

$$\partial h(x) = \{x^*|\langle x^*, z - x\rangle \leqslant h(z) - h(x), \text{ for all } z\}. \tag{A.1}$$

The domain of ∂h is the set dom $\partial h = \{x|\partial h(x) \neq \emptyset\}$. The elements of the subdifferential are *subgradients*. If h is differentiable, then the subdifferential contains only the gradient, that is,

$$\partial h(x) = \{\nabla h(x)\}. \tag{A.2}$$

The *conjugate function* associated with h is the function

$$h^*(a) = \sup_x(\langle a, x\rangle - h(x)). \tag{A.3}$$

A.2. Essential smoothness and essential strict convexity

Following [22] we say that a closed proper convex function h is *essentially smooth* if int D, the interior of D, is not empty, h is differentiable on int D, and $x^n \in$ int D, with $x^n \to x$, with x in bdry D, the boundary of D, implies that $||\nabla h(x^n)|| \to +\infty$.

A closed proper convex function h is *essentially strictly convex* if h is strictly convex on every convex subset of dom ∂h.

The closed proper convex function h is essentially smooth if and only if the subdifferential $\partial h(x)$ is empty for $x \in$ bd$D$ and is $\{\nabla h(x)\}$ for $x \in$ int D (so h is differentiable on int D) if and only if the function h^* is essentially strictly convex.

A closed proper convex function h is said to be a *Legendre function* if it is both essentially smooth and essentialy strictly convex. So h is Legendre if and only if its conjugate function is Legendre, in which case the gradient operator ∇h is a topological isomorphism with ∇h^* as its inverse. The gradient operator ∇h maps int dom h onto int dom h^*. If int dom $h^* = R^J$ then the range of ∇h is R^J and the equation $\nabla h(x) = y$ can be solved for every $y \in R^J$. In order for int dom $h^* = R^J$ it is necessary and sufficient that the Legendre function h be *super-coercive*, that is,

$$\lim_{||x||\to+\infty} \frac{h(x)}{||x||} = +\infty.$$

If the essential domain of h is bounded, then h is super-coercive and its gradient operator is a mapping onto the space R^J.

A.3. Bregman projections onto closed convex sets

Let K be a nonempty closed convex set with $K \cap$ int $D \neq \emptyset$. Pick $z \in$ int D. The *Bregman projection* of z onto K, with respect to h, is

$$P_K^h(z) = \text{argmin}_{x\in K\cap D} D_h(x, z).$$

If h is essentially strictly convex, then $P_K^h(z)$ exists. If h is strictly convex on D then $P_K^h(z)$ is unique. If h is Legendre, then $P_K^f(z)$ is uniquely defined and is in int D; this last condition is sometimes called *zone consistency*.

Example. Let $J = 2$ and $h(x)$ be the function that is equal to one-half the norm squared on D, the nonnegative quadrant, $+\infty$ elsewhere. Let K be the set $K = \{(x_1, x_2)|x_1 + x_2 = 1\}$. The Bregman projection of $(2, 1)$ onto K is $(1, 0)$, which is not in int D. The function h is not essentially smooth, although it is essentially strictly convex. Its conjugate is the function h^* that is equal to one-half the norm squared on D and equal to zero elsewhere; it is essentially smooth, but not essentially strictly convex.

If h is Legendre, then $P_K^h(z)$ is the unique member of $K \cap$ int D satisfying the inequality

$$\left\langle \nabla h\left(P_K^h(z)\right) - \nabla h(z),\, P_K^h(z) - c\right\rangle \geqslant 0, \tag{A.4}$$

for all $c \in K$. From this we obtain the *Bregman Inequality*:

$$D_h(c, z) \geqslant D_h\left(c,\, P_K^h(z)\right) + D_h\left(P_K^h(z), z\right), \tag{A.5}$$

for all $c \in K$.

A.4. Bregman–Legendre functions

Following Bauschke and Borwein [4], we say that a Legendre function h is a *Bregman–Legendre* function if the following properties hold:

B1: for x in D and any $a > 0$ the set $\{z|D_h(x, z) \leqslant a\}$ is bounded.

B2: if x is in D but not in int D, for each positive integer n, y^n is in int D with $y^n \to y \in$ bd D and if $\{D_h(x, y^n)\}$ remains bounded, then $D_h(y, y^n) \to 0$, so that $y \in D$.

B3: if x^n and y^n are in int D, with $x^n \to x$ and $y^n \to y$, where x and y are in D but not in int D, and if $D_h(x^n, y^n) \to 0$ then $x = y$.

A.5. Useful results about Bregman–Legendre functions

The following results are proved in somewhat more generality in [4].

R1: If $y^n \in$ int dom h and $y^n \to y \in$ int dom h, then $D_h(y, y^n) \to 0$.

R2: If x and $y^n \in$ int dom h and $y^n \to y \in$ bd dom h, then $D_h(x, y^n) \to +\infty$.

R3: If $x^n \in D$, $x^n \to x \in D$, $y^n \in$ int D, $y^n \to y \in D$, $\{x, y\} \cap$ int $D \neq \emptyset$ and $D_h(x^n, y^n) \to 0$, then $x = y$ and $y \in$ int D.

R4: If x and y are in D, but are not in int D, $y^n \in$ int D, $y^n \to y$ and $D_h(x, y^n) \to 0$, then $x = y$.

As a consequence of these results we have the following.

R5: If $\{D_h(x, y^n)\} \to 0$, for $y^n \in$ int D and $x \in R^J$, then $\{y^n\} \to x$.

Proof of R5. Since $\{D_h(x, y^n)\}$ is eventually finite, we have $x \in D$. By property B1 above it follows that the sequence $\{y^n\}$ is bounded; without loss of generality, we assume that $\{y^n\} \to y$, for some $y \in \overline{D}$. If x is in int D, then, by result R2 above, we know that y is also in int D. Applying result R3, with $x^n = x$, for all n, we conclude that $x = y$. If, on the other hand, x is in D, but not in int D, then y is in D, by result R2. There are two cases to consider: (1) y is in int D; (2) y is not in int D. In case (1) we have $D_h(x, y^n) \to D_h(x, y) = 0$, from which it follows that $x = y$. In case (2) we apply result R4 to conclude that $x = y$. □

References

[1] Attouch H 1984 *Variational Convergence for Functions and Operators* (Boston: Pitman Advanced Publishing Program)

[2] Attouch H and Wets R 1989 Epigraphical analysis *Ann. Inst. Poincare: Anal. Nonlineaire* **6**

[3] Auslander A and Teboulle M 2006 Interior gradient and proximal methods for convex and conic optimization *SIAM J. Optim.* **16** 697–725

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

[5] Bauschke H, Combettes P and Noll D 2006 Joint minimization with alternating Bregman proximity operators *Pac. J. Optim.* **2** 401–24

[6] Byrne C 1993 Iterative image reconstruction algorithms based on cross-entropy minimization *IEEE Trans. Image Process.* **2** 96–103

[7] Byrne C 1998 Iterative algorithms for deblurring and deconvolution with constraints *Inverse Problems* **14** 1455–67

[8] Byrne C 2001 Bregman–Legendre multidistance projection algorithms for convex feasibility and optimization *Inherently Parallel Algorithms in Feasibility and Optimization and their Applications* ed D Butnariu, Y Censor and S Reich (*Studies in Computational Mathematics* vol 8) (Amsterdam: Elsevier) pp 87–100

[9] Byrne C 2002 Iterative oblique projection onto convex sets and the split feasibility problem *Inverse Problems* **18** 441–53
[10] Byrne C 2004 A unified treatment of some iterative algorithms in signal processing and image reconstruction *Inverse Problems* **20** 103–20
[11] Byrne C 2005 Choosing parameters in block-iterative or ordered-subset reconstruction algorithms *IEEE Trans. Image Process.* **14** 321–7
[12] Byrne C 2005 *Signal Processing: A Mathematical Approach* (Wellesley, MA: AK Peters)
[13] Censor Y and Zenios S A 1992 Proximal minimization algorithm with *D*-functions *J. Optim. Theory Appl.* **73** 451–64
[14] Combettes P and Wajs V 2005 Signal recovery by proximal forward–backward splitting *Multiscale Model. Simul.* **4** 1168–200
[15] Fiacco A and McCormick G 1990 Nonlinear programming: sequential unconstrained minimization techniques *SIAM Classics in Mathematics* (Philadelphia, PA: SIAM) (reissue)
[16] Geman S and Geman D 1984 Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images *IEEE Trans. Pattern Anal. Mach. Intell.* **6** 721–41
[17] 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.* **6** 106–14
[18] Lange K and Carson R 1984 EM reconstruction algorithms for emission and transmission tomography *J. Comput. Assist. Tomogr.* **8** 306–16
[19] Moreau J-J 1963 Propriétés des applications prox *C. R. Acad. Sci. Paris* A **256** 1069–71
[20] Nesterov Y and Nemirovski A 1994 Interior-point polynomial algorithms in convex programming *SIAM Studies in Applied Mathematics* (Philadelphia, PA: SIAM)
[21] Renegar J 2001 *A Mathematical View of Interior-Point Methods in Convex Optimizations (MPS-SIAM Series on Optimization)* (Philadelphia, PA: SIAM)
[22] Rockafellar R 1970 *Convex Analysis* (Princeton, NJ: Princeton University Press)
[23] Rockafellar R and Wets R J B 1998 *Variational Analysis* (New York: Springer)
[24] Teboulle M 1992 Entropic proximal mappings with applications to nonlinear programming *Math. Oper. Res.* **17** 670–90
[25] Vardi Y, Shepp L A and Kaufman L 1985 A statistical model for positron emission tomography *J. Am. Stat. Assoc.* **80** 8–20