Iterative Convex Optimization Algorithms; Part Two: Without the Baillon–Haddad Theorem

Charles Byrne
(Charles_Byrne@uml.edu)
http://faculty.uml.edu/cbyrne/cbyrne.html
Department of Mathematical Sciences
University of Massachusetts Lowell
Lowell, MA 01854, USA

February 25, 2015

Avoiding the Baillon–Haddad Theorem

I first became interested in iterative algorithms as applied to medical image reconstruction, where the vectors involved are nonnegative and the cross-entropy distance plays an important role. I became aware of the BH Theorem only about ten years ago. Until this year there was no elementary proof of the Baillon–Haddad Theorem in the published literature. For these reasons, my investigation of convergence of certain iterative optimization algorithms involved methods that do not rely on the BH Theorem. The results of this investigation provide the subject matter of this talk.

The Basic Problem

The basic problem is to minimize $f : X \to \mathbb{R}$, over x in $C \subseteq X$, where, initially, X is an arbitrary nonempty set. Initially, X has no structure for two reasons:

1. we can do quite a bit without structure, and
2. we want to consider several different structures.

Barrier- and Penalty-Function Methods

Barrier-function and penalty-function methods are the best known of the sequential unconstrained minimization techniques discussed by Fiacco and McCormack. In barrier-function methods the **barrier function** $b(x)$ is chosen to be positive for x in C and infinite for x outside C. We then minimize

$$B_k(x) = f(x) + \frac{1}{k}b(x)$$

to get x^k. In penalty-function methods we select the **penalty function** $p(x) > 0$ for x not in C and $p(x) = 0$ for x in C. Then we minimize

$$P_k(x) = f(x) + kp(x)$$

to get x^k. For penalty-function methods we need topology on X and f and p continuous.

An Example of Barrier-Function Methods

Minimize $f(x) = x_1^2 + x_2^2$, subject to $x_1 + x_2 \geq 1$. Let

$$b(x) = -\log(x_1 + x_2 - 1),$$

for $x_1 + x_2 > 1$, and $b(x) = +\infty$, otherwise. For each k we minimize

$$B_k(x) = f(x) + \frac{1}{k}b(x)$$

to get

$$x_1^k = x_2^k = \frac{1}{4} + \frac{1}{4}\sqrt{1 + \frac{4}{k}}.$$

As $k \to +\infty$, $\{x^k\} \to (\frac{1}{2}, \frac{1}{2})$.

An Example of Penalty-Function Methods

We want to minimize the function $f(x) = (x+1)^2$, subject to $x \geq 0$. Let $p(x) = 0$, for $x \geq 0$, and $p(x) = x^2$, for $x < 0$. Then we minimize

$$P_k(x) = f(x) + kp(x)$$

to get $x^k = \frac{-1}{k+1}$. As $k \to +\infty$, $\{x^k\} \to 0$. Note that the sequence $\{f(x^k) = \frac{k}{k+1}\}$ is increasing; but each x^k is outside C.

Penalty Methods as Barrier Methods

Let f be bounded below, so that, without loss of generality, we may assume $f : \mathbb{R}^J \to \mathbb{R}_+$, $p : \mathbb{R}^J \to [0, +\infty)$, and $C = \{x | p(x) = 0\}$. For each k we minimize

$$P_k(x) = f(x) + kp(x)$$

to get x^k. Equivalently, we can minimize

$$p(x) + \frac{1}{k}f(x),$$

which has the form of a barrier-function method.

$\{f(x^k)\} \downarrow \inf_{x \in C} f(x)$ for Barrier-Function Methods

From $B_k(x^{k-1}) \geq B_k(x^k)$ and $B_{k-1}(x^k) \geq B_{k-1}(x^{k-1})$, for $k = 2, 3, ...$, it follows easily that

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

Suppose that $\{f(x^k)\} \downarrow \beta^* > \beta = \inf_{x \in C} f(x)$. Then there is $z \in C$ with

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

for all k. Then

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

for all k. But the sequence $\{\frac{1}{k}(b(z) - b(x^k))\}$ converges to zero, which contradicts the assumption that $\beta^* > \beta$.

Applying the Penalty-as-Barrier Method

For each k we minimize

$$P_k(x) = f(x) + kp(x)$$

to get x^k. Since penalty-function methods can be reformulated as barrier-function methods, we can show that

$$\{P_k(x^k)\} \uparrow \gamma \leq \beta = \inf_{x \in C} f(x),$$

$$\{p(x^k)\} \downarrow 0,$$

and

$$\{f(x^k)\} \uparrow \gamma^* \leq \gamma \leq \beta.$$

For more we need X to be a complete metric space, f and p to be continuous, and f to have compact level sets.

The Basic Approach

We study iterative algorithms in which, having found x^{k-1}, we minimize

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

over x in C to get x^k. We are particularly interested in how to restrict the g_k to impose desirable properties on the sequence $\{x^k\}$, such as

1. the sequence $\{f(x^k)\} \downarrow \beta^* \geq -\infty$ is nonincreasing;
2. the sequence $\{f(x^k)\}$ converges to $\beta = \inf_{x \in C} f(x)$, so $\beta^* = \beta$; and
3. the sequence $\{x^k\}$ converges to $x^* \in C$ with $f(x^*) \leq f(x)$, for all $x \in C$.

Auxiliary-Function Methods

Definition 1. The functions $g_k(x)$ are **auxiliary functions** (AF) if they have the properties $g_k(x) \geq 0$ for all $x \in C$, and $g_k(x^{k-1}) = 0$.

Lemma 2. *If the sequence $\{x^k\}$ is generated by an AF method, then the sequence $\{f(x^k)\}$ is nonincreasing.*

Proof: We have

$$G_k(x^{k-1}) = f(x^{k-1}) + g_k(x^{k-1}) = f(x^{k-1})$$
$$\geq G_k(x^k) = f(x^k) + g_k(x^k) \geq f(x^k).$$

Barrier-Function Methods as AF Methods

The iterative step in barrier-function methods is to minimize

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

to get x^k. Equivalently, we can minimize

$$kf(x) + b(x) = f(x) + (k-1)f(x) + b(x),$$

or even

$$f(x) + [(k-1)f(x) + b(x)] - [(k-1)f(x^{k-1}) + b(x^{k-1})]$$
$$= f(x) + g_k(x) = G_k(x).$$

It is easy to show that $G_k(x) - G_k(x^k) = g_{k+1}(x)$, so barrier-function methods satisfy the SUMMA Inequality, to be defined later.

Generalized Proximal Minimization

A large subclass of AF methods is the class of **generalized proximal minimization algorithms**. For each k let $d_k : X \times X \to \mathbb{R}_+$ be a "distance", so that $d_k(x, x) = 0$. Then minimize

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

to get x^k.

Proximal minimization algorithms (PMA) require that $d_k = d$ for all k, so for PMA we minimize

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

to get x^k.

Proximal Minimization Algorithms

There are a variety of PMA algorithms. For example,

- the EM algorithm and cross-entropy minimization;
- the CQ algorithm for the split feasibility problem;
- projected gradient descent algorithms;
- majorization minimization (MM) in statistics (K. Lange, et al.);
- the optimization method of Auslender and Teboulle;
- proximal minimization with Bregman distances (PMAB) (Censor and Zenios);
- forward-backward splitting (FBS) (Combettes and Wajs).

MM=PMA

The *majorization minimization* (MM) method in statistics, also called *optimization transfer*, is the following. Assume that there is a function $g(x|y) \geq f(x)$, for all x and y, with $g(y|y) = f(y)$. Then, for each k, minimize $g(x|x^{k-1})$ to get x^k.

The MM methods and the PMA methods are equivalent; given $g(x|y)$, define $d(x,y) \doteq g(x|y) - f(x)$ and given $d(x,y)$, define $g(x|y) \doteq f(x) + d(x,y)$.

The SUMMA Inequality

In order to have the AF sequence $\{f(x^k)\}$ converge to $\beta = \inf\{f(x)|x \in C\}$ we need to impose an additional property on the $g_k(x)$.

Definition 3. An AF method is in the **SUMMA class** if the **SUMMA Inequality** holds:

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

for all $x \in C$.

So barrier-function methods are in the SUMMA class.

The SUMMA Class

The SUMMA class of algorithms may seem to be quite limited, but, as we shall see, that is far from the case. Many well known iterative methods, including barrier-function methods, proximal minimization using Bregman distances, the SMART for cross-entropy minimization, and alternating minimization with the five-point property, are in the SUMMA class or can be reformulated to be in this class (CB, 2008).

The SUMMA Theorem

We have the following theorem.

Theorem 4. *If the auxiliary functions g_k satisfy the SUMMA Inequality, then the sequence $\{f(x^k)\}$ converges to $\beta = \inf_{x \in C} f(x)$.*

Proof: If not, then there is $z \in C$ and β^* such that

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

From

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

it follows that $\{g_k(z)\}$ is a decreasing sequence of nonnegative terms whose successive differences remain bounded away from zero, which is a contradiction.

The PMAB

Proximal minimization algorithms using Bregman distances, called here PMAB, form an important subclass of the SUMMA class. Let $f : \mathbb{R}^J \to \mathbb{R}$ be a convex differentiable function. Once again, our objective is to minimize $f(x)$ over x in the nonempty closed convex set C. Assume that $f(x)$ attains its minimum value on C at $\hat{x} \in C$.

Let h be another convex differentiable function, which may or may not involve C. If it does, then its effective domain is $C = \{x | h(x) < +\infty\}$, and h is differentiable on the nonempty open convex set int(C). If not, we redefine $f(x)$ by $f(x) = +\infty$, for x outside C. Then the minimization is automatically over $x \in C$.

The PMAB Iteration

At the kth step of a **PMAB** we minimize

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

to get x^k. The **Bregman distance**

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

is sometimes called a **proximity function**. The function $g_k(x) = D_h(x, x^{k-1})$ is nonnegative for $x \in C$ and $g_k(x^{k-1}) = 0$. Then

$$G_k(x) - G_k(x^k) = D_f(x, x^k) + D_h(x, x^k) \geq D_h(x, x^k) = g_{k+1}(x) \geq 0,$$

so all PMAB are in the SUMMA class.

PMAB: The Non-differentiable Case

In the PMAB we can remove the requirement that h be differentiable, and just assume that h is convex. Then, to show that PMAB is in the SUMMA class we need the *subdifferential* of h at x,

$$\partial h(x) = \{u | h(y) - h(x) \geq \langle u, y - x \rangle \, , \text{for all} \, y\}.$$

A function $h : \mathbb{R}^J \to \mathbb{R}$ is convex if and only if $\partial h(x)$ is not empty, for every x. Note that $x = z$ minimizes $h(x)$ if and only if $0 \in \partial h(z)$. If f and h are convex and $f + h = a$ is differentiable, then both f and h are differentiable.

Computational Difficulty with PMAB

To minimize $G_k(x) = f(x) + D_h(x, x^{k-1})$ we must solve

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

for x^k. This can be difficult. Suppose we select $a(x)$ so that

$$h(x) \doteq a(x) - f(x)$$

is convex and differentiable, and

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

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

is easily solved. The *projected gradient descent* (PGD) algorithm is a good example of this *modified PMAB* method.

Orthogonal Projection

Let $C \subseteq \mathcal{H}$ be a nonempty, closed, convex set. For every x in $\mathcal{H}$ there is a unique member of C closest to x, called the *orthogonal projection* of x onto C and denoted $P_C x$.

Theorem 5. *A vector z in C is $P_C x$ if and only if*

$$\langle z - x, c - z \rangle \geq 0,$$

for all c in C.

Minimizing $f(x)$ over $x \in C$

Theorem 6. *Let $f : \mathcal{H} \to \mathbb{R}$ be convex and differentiable. A vector $z \in C$ minimizes $f(x)$ over $x \in C$ if and only if*

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

for all $c \in C$

Then, for all $\gamma > 0$,

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

so that

$$z = P_C(z - \gamma \nabla f(z)).$$

Projected Gradient Descent (PGD) Algorithms

The problem now is to minimize $f : \mathbb{R}^J \to \mathbb{R}$, over the closed, nonempty convex set C, where f is convex and differentiable on $\mathbb{R}^J$ and ∇f is L-Lipschitz continuous;

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|.$$

For $0 < \gamma < \frac{1}{L}$ let

$$a(x) = \frac{1}{2\gamma}\|x\|_2^2;$$

then the function $h(x) = a(x) - f(x)$ is convex. At the kth step we use the modified PMAB approach to obtain x^k by minimizing

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

over $x \in C$.

The PGD Continued

To be the minimizer of $G_k(x)$ over $x \in C$, the vector x^k must satisfy the inequalities

$$\langle \nabla G_k(x^k), c - x^k \rangle \geq 0,$$

for all c in C. Therefore, the solution x^k is in C and satisfies the inequality

$$\langle x^k - (x^{k-1} - \gamma \nabla f(x^{k-1})), c - x^k \rangle \geq 0,$$

for all $c \in C$. It follows then that

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

Comments on the Modified PMAB

Note that the auxiliary function for the PGD,

$$g_k(x) = \frac{1}{2\gamma}\|x - x^{k-1}\|_2^2 - D_f(x, x^{k-1}) = D_h(x, x^{k-1}),$$

is unrelated to the set C, so is used here not to incorporate the constraint, but to provide a closed-form iterative scheme. When $C = \mathbb{R}^J$ we have no constraint and the problem is simply to minimize f. Then the iterative algorithm becomes

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

this is the **gradient descent algorithm**.

The Projected Landweber Algorithm

The **Landweber** (LW) and **projected Landweber** (PLW) algorithms are special cases of PGD. The objective now is to minimize the function

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

over $x \in \mathbb{R}^J$ or $x \in C$, where A is a real matrix. Then

$$\nabla f(x) = A^T(Ax - b)$$

is L-Lipschitz continuous for $L = \rho(A^T A)$ and

$$D_f(x, z) = \frac{1}{2}\|Ax - Az\|_2^2.$$

The PLW Iteration

We let

$$a(x) = \frac{1}{2\gamma}\|x\|_2^2,$$

where $0 < \gamma < \frac{1}{L}$, so that the function $h(x) = a(x) - f(x)$ is convex. At the kth step of the PLW we minimize

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

over $x \in C$ to get the **projected Landweber** iteration

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

in the case of $C = \mathbb{R}^J$ we get the **Landweber algorithm**.

Minimize $f(x) = \frac{1}{2}\|Ax - b\|^2$ **over** C

For all $c \in C$ we have

$$\langle \nabla G_k(x^k), c - x^k \rangle \geq 0,$$

where

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

Therefore,

$$\langle x^k - (x^{k-1} - \gamma A^T(Ax^{k-1} - b)), c - x^k \rangle \geq 0,$$

or

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

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

For each $k = 1, 2, ...$, minimize

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

to get the **Landweber** iteration

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

We select $0 < \gamma < \frac{1}{\rho(A^TA)}$. The added function

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

serves to give x^k in closed form.

The Kullback-Leibler Distance

For $a > 0$ and $b > 0$ define

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

with $KL(0, b) = b$ and $KL(a, 0) = +\infty$. Extend component-wise to $KL(x, z)$, for any nonnegative vectors x and z. Then

$$KL(x, z) = \sum_{j=1}^{J} x_j \log \frac{x_j}{z_j} + z_j - x_j \geq 0,$$

and $KL(x, z) = 0$ if and only if $x = z$.

Minimize $f(x) = KL(Px, y)$ **over** $x \geq 0$

Now y is a positive vector, P an I by J matrix with nonnegative entries, whose columns sum to one. For $k = 1, 2, ...$, we minimize

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

to get

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

Then $\{x^k\}$ converges to the minimizer of $KL(Px, y)$ for which $KL(x, x^0)$ is minimized.

Proximal Minimization of Auslender and Teboulle

Auslender and Teboulle take C to be a closed, nonempty, convex subset of $\mathbb{R}^J$, with interior U. At the kth step of this AT method one minimizes a function

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

to get x^k. Their distance $d(x, y)$ is defined for x and y in U, and the gradient with respect to the first variable, denoted $\nabla_1 d(x, y)$, is assumed to exist. The distance $d(x, y)$ is not assumed to be a Bregman distance.

Associated Induced Proximal Distance

Instead, they assume that the distance d has an associated *induced proximal distance* $H(a,b) \geq 0$, finite for a and b in U, with $H(a,a) = 0$ and

$$\langle \nabla_1 d(b,a), c-b \rangle \leq H(c,a) - H(c,b),$$

for all c in U. For such algorithms we have $\beta^* = \beta$, although this method is not in the SUMMA class. We want to extend the SUMMA class to include the AT algorithms of Auslender and Teboulle.

The SUMMA2 Class

The AT proximal minimization method of Auslender and Teboulle is not in the SUMMA class, although $\beta^* = \beta$. A consequence of the SUMMA Inequality is

$$g_k(x) - g_{k+1}(x) \geq f(x^k) - f(x), \text{for all } x \in C,$$

and for the AT methods we have

$$H(x, x^{k-1}) - H(x, x^k) \geq f(x^k) - f(x), \text{for all } x \in C,$$

We say that an AF algorithm is in the SUMMA2 class (CB, 2015) if, for any sequence $\{x^k\}$ generated by the algorithm, with $x^k \in C$, there are functions $h_k : X \to [0, +\infty]$, finite-valued on C, such that

$$h_k(x) - h_{k+1}(x) \geq f(x^k) - f(x), \text{for all } x \in C.$$

With $h_k(x) = H(x, x^{k-1})$ the AT algorithms are in SUMMA2.

The SUMMA2 Theorem

We have the following theorem.

Theorem 7. *If $\{x^k\}$ is generated by a SUMMA2 algorithm then $\beta^* = \beta$.*

Proof: If not, then there is $z \in C$ such that

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

From

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

it follows that $\{h_k(z)\}$ is a decreasing sequence of nonnegative terms whose successive differences remain bounded away from zero, which is a contradiction.

Auslender-Teboulle algorithms are in SUMMA2

Since x^k minimizes $f(x) + d(x, x^{k-1})$, it follows that

$$0 \in \partial f(x^k) + \nabla_1 d(x^k, x^{k-1}),$$

so that

$$-\nabla_1 d(x^k, x^{k-1}) \in \partial f(x^k).$$

We then have

$$f(x^k) - f(x) \leq \langle \nabla_1 d(x^k, x^{k-1}), x - x^k \rangle.$$

Using the associated induced proximal distance H, we obtain

$$H(x, x^{k-1}) - H(x, x^k) \geq f(x^k) - f(x).$$

So $h_k(x) = H_k(x, x^{k-1})$ works.

The Forward-Backward Splitting Algorithm

The **forward-backward splitting** (FBS) methods form a broad class of PMAB algorithms. We want to minimize the function

$$f(x) = f_1(x) + f_2(x),$$

where both functions are convex and $f_2(x)$ is differentiable with L-Lipschitz continuous gradient. At the kth step of the FBS algorithm we obtain x^k by minimizing

$$G_k(x) = f_1(x) + f_2(x) + \frac{1}{2\gamma}\|x - x^{k-1}\|_2^2 - D_{f_2}(x, x^{k-1}),$$

over all $x \in \mathbb{R}^J$, where $0 < \gamma \leq \frac{1}{L}$.

Moreau's Proximity Operators

The **Moreau envelope** of index $\gamma > 0$ of the convex function $f : \mathbb{R}^J \to \mathbb{R}$ is the convex function

$$g(x) = \inf\{f(y) + \frac{1}{2\gamma}||x - y||_2^2\}.$$

The infimum is attained at a unique $y = \text{prox}_{\gamma f}(x)$. The subdifferential of f at x is the set

$$\partial f(x) = \{u | \langle u, z - x \rangle \leq f(z) - f(x)\}$$

for all z and x. Note that $f(z) \leq f(x)$, for all x, if and only if $0 \in \partial f(z)$.

Properties of prox$_{\gamma f}$

The proximity operators $\text{prox}_{\gamma f}(\cdot)$ has several useful properties.

- These operators are firmly nonexpansive.
- The vector x is $\text{prox}_f(z)$ if and only if $z - x \in \partial f(x)$.
- For $f(x) = \iota_C(x)$, which is zero for $x \in C$ and infinity elsewhere, we have $\text{prox}_{\iota_C}(x) = P_C(x)$.

Combettes and Wajs noticed that x minimizes $f_1(x) + f_2(x)$ over $x \in \mathbb{R}^J$ if and only if

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

which suggested the FBS iteration. They did not consider the FBS as a PMAB algorithm.

The FBS Iteration

Let $f : \mathbb{R}^J \to \mathbb{R}$ be convex, with $f = f_1 + f_2$, both convex, f_2 differentiable, and ∇f_2 L-Lipschitz continuous. The iterative step of the FBS algorithm is

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

Convergence of the sequence $\{x^k\}$ to a solution can be established, if γ is chosen to lie within the interval $(0, 1/L]$.

The CQ Algorithm

Let A be a real I by J matrix, $C \subseteq \mathbb{R}^J$, and $Q \subseteq \mathbb{R}^I$, both closed convex sets. The **split feasibility problem** (SFP) is to find x in C such that Ax is in Q. The function

$$f_2(x) = \frac{1}{2}\|P_Q Ax - Ax\|_2^2$$

is convex, differentiable and $\nabla f_2 = A^T(I - P_Q)Ax$ is L-Lipschitz for $L = \rho(A^T A)$. We want to minimize the function $f(x) = \iota_C(x) + f_2(x)$. The CQ algorithm (CB,2002) has the iterative step

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

The sequence converges to a solution whenever f_2 has a minimum on the set C, for $0 < \gamma \leq 1/L$.

Intensity Modulated Radiation Therapy

Yair Censor and colleagues modified the CQ algorithm to obtain efficient algorithms for designing protocols for **intensity modulated radiation therapy (IMRT)**.

- Censor, Y., Bortfeld, T., Martin, B., and Trofimov, A. "A unified approach for inversion problems in intensity-modulated radiation therapy." *Physics in Medicine and Biology* 51 (2006), 2353-2365.
- Censor, Y., Elfving, T., Kopf, N., and Bortfeld, T. (2005) "The multiple-sets split feasibility problem and its application for inverse problems." *Inverse Problems*, **21** , pp. 2071-2084.

Alternating Minimization (AM)

We turn now to the apparently unrelated alternating minimization (AM) method of Csiszár and Tusnády.

Suppose that X and Y are arbitrary nonempty sets and the function $\Theta(x, y)$ satisfies $-\infty < \Theta(x, y) \leq +\infty$, for each $x \in X$ and $y \in Y$. The objective is to generate a sequence $\{(x^k, y^k)\}$ such that

$$\Theta(x^k, y^k) \downarrow \beta = \inf_{x \in X, y \in Y} \Theta(x, y).$$

The AM Iteration

The general AM method of Csiszár and Tusnády proceeds in two steps:

we begin with some y^0, and, having found y^k, we

1. minimize $\Theta(x, y^k)$ over $x \in X$ to get $x = x^{k+1}$, and then

2. minimize $\Theta(x^{k+1}, y)$ over $y \in Y$ to get $y = y^{k+1}$.

The Five-Point Property for AM

The **five-point property** is the following: for all $x \in X$ and $y \in Y$ and $k = 1, 2, ...$

$$\Theta(x, y) + \Theta(x, y^{k-1}) \geq \Theta(x, y^k) + \Theta(x^k, y^{k-1}),$$

or

$$\Theta(x, y^{k-1}) - \Theta(x^k, y^{k-1}) \geq \Theta(x, y^k) - \Theta(x, y).$$

The Main Theorem for AM

Theorem 8. *If the five-point property holds then*

$$\Theta(x^k, y^k) \downarrow \beta = \inf_{x \in X, y \in Y} \Theta(x, y).$$

Reformulating AM as AF

For each x in the set X, define $y(x)$ in Y as a member of Y for which $\Theta(x, y(x)) \leq \Theta(x, y)$, for all $y \in Y$. Let

$$f(x) = \Theta(x, y(x)).$$

Now we want

$$f(x^k) \downarrow \beta = \inf_{x \in X} f(x).$$

AM as SUMMA

At the kth step of AM we minimize

$$G_k(x) = \Theta(x, y^{k-1}) = \Theta(x, y(x)) + \Big(\Theta(x, y^{k-1}) - \Theta(x, y(x))\Big)$$

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

to get x^k. Then the five-point property is precisely the SUMMA Inequality:

$$G_k(x) - G_k(x^k) = \Theta(x, y^{k-1}) - \Theta(x^k, y^{k-1})$$

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

Convergence of the PGD Algorithm

A relatively simple calculation shows that, for all $x \in C$,

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

$$\frac{1}{\gamma}\langle x^k - (x^{k-1} - \gamma\nabla f(x^{k-1})), x - x^k\rangle \geq \frac{1}{2\gamma}\|x - x^k\|_2^2.$$

Therefore, for all $x \in C$, we have

$$G_k(x) - G_k(x^k) \geq \frac{1}{2\gamma}\|x - x^k\|_2^2 - D_f(x, x^k) = g_{k+1}(x).$$

Convergence Proof Continued

Now let $\hat{x}$ minimize $f(x)$ over all x. Then

$$G_k(\hat{x}) - G_k(x^k) = f(\hat{x}) + g_k(\hat{x}) - f(x^k) - g_k(x^k)$$

$$\leq f(\hat{x}) + G_{k-1}(\hat{x}) - G_{k-1}(x^{k-1}) - f(x^k) - g_k(x^k),$$

so that

$$\Big(G_{k-1}(\hat{x}) - G_{k-1}(x^{k-1})\Big) - \Big(G_k(\hat{x}) - G_k(x^k)\Big)$$

$$\geq f(x^k) - f(\hat{x}) + g_k(x^k) \geq 0.$$

Therefore, the sequence $\{G_k(\hat{x}) - G_k(x^k)\}$ is decreasing and the sequences $\{g_k(x^k)\}$ and $\{f(x^k) - f(\hat{x})\}$ converge to zero.

More Convergence Proof Continued

From

$$G_k(\hat{x}) - G_k(x^k) \geq \frac{1}{2\gamma}\|\hat{x} - x^k\|_2^2,$$

it follows that the sequence $\{x^k\}$ is bounded. Therefore, we may select a subsequence $\{x^{k_n}\}$ converging to some x^{**}, with $\{x^{k_n-1}\}$ converging to some x^*, and therefore $f(x^*) = f(x^{**}) = f(\hat{x})$. Replacing the generic $\hat{x}$ with x^{**}, we find that $\{G_k(x^{**}) - G_k(x^k)\}$ is decreasing to zero. We conclude that the sequence $\{\|x^* - x^k\|_2^2\}$ converges to zero, and so $\{x^k\}$ converges to x^*.

The End

THE END

References

- 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.

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

- Byrne, C. (2001) "Bregman-Legendre multi-distance projection algorithms for convex feasibility and optimization." in Butnariu, et al., pp. 87–100.

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

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

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

- 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)**, pp. 554–566.

- 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.

- Byrne, C. (2014) "On a generalized Baillon–Haddad Theorem for convex functions on Hilbert space." to appear in the *Journal of Convex Analysis*.

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

- Byrne, C. (2015) "Auxiliary-function methods in iterative optimization." submitted.

- Chi, E., Zhou, H., and Lange, K. (2014) "Distance Majorization and Its Applications." Mathematical Programming, **146 (1-2)**, pp. 409–436.

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

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

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

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

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

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

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

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

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