Alternating Minimization, Optimization Transfer and Proximal Minimization Are Equivalent (9/16/15 draft)

Charles L. Byrne*

September 16, 2015

Abstract

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

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

it is clear that MM is equivalent to PMA. Alternating minimization (AM) methods appear to be more general, but AM is equivalent to PMA and to MM.

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

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

for all $x \in X$ and $y \in Y$. For each k we minimize $\Phi(x, y^{k-1})$ to get x^{k-1} and then minimize $\Phi(x^{k-1}, y)$ to get y^k. For each $x \in X$, let $y(x) \in Y$ be such that $\Phi(x, y) \geq \Phi(x, y(x))$, for all $y \in Y$; then $y^k = y(x^{k-1})$. Minimizing $\Phi(x, y)$ over all $x \in X$ and $y \in Y$ is equivalent to minimizing $f(x) \doteq \Phi(x, y(x))$ over all $x \in X$. With $d(x, x') = \Phi(x, y(x')) - \Phi(x, y(x))$, minimizing $\Phi(x, y^k)$ is equivalent to minimizing $f(x) + d(x, x^{k-1})$. Therefore, all AM algorithms are instances of PMA, and therefore, of MM.

*Charles_Byrne@uml.edu, Department of Mathematical Sciences, University of Massachusetts Lowell, Lowell, MA 01854

1 Auxiliary-Function Methods in Optimization

Let $f : X \to \mathbb{R}$, where X is an arbitrary nonempty set. In applications the set X will have additional structure, but not always that of a Euclidean space; for that reason, it is convenient to impose no structure at the outset. An iterative procedure for minimizing $f(x)$ over $x \in X$ is called an *auxiliary-function* (AF) algorithm [4, 7] if, at each step, we minimize

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

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

1.1 The SUMMA Inequality

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

$$G_k(x) - G_k(x^k) \geq g_{k+1}(x). \tag{1.2}$$

One consequence of the SUMMA Inequality is

$$g_k(x) + f(x) \geq g_{k+1}(x) + f(x^k), \tag{1.3}$$

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

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

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

There are many iterative algorithms that satisfy the SUMMA Inequality [4], and are therefore in the SUMMA class. However, some important methods that are not in this class still have $\beta^* = \beta$; one example is the proximal minimization method of Auslender and Teboulle [2]. This suggests that the SUMMA class, large as it is, is still unnecessarily restrictive. This leads us to the definition of the SUMMA2 class.

1.2 The SUMMA2 Class

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

such that, for all $x \in X$, we have

$$h_k(x) + f(x) \geq h_{k+1}(x) + f(x^k). \tag{1.4}$$

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

2 PMA is MM

In proximal minimization algorithms (PMA) we minimize

$$f(x) + d(x, x^{k-1}) \tag{2.1}$$

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

In [8] the authors review the use, in statistics, of "majorization minimization" (MM), also called "optimization transfer". In numerous papers [10, 1] Jeff Fessler and his colleagues use the terminology "surrogate-function minimization" to describe optimization transfer. The objective is to minimize $f : X \to \mathbb{R}$. In MM methods a second "majorizing" function $g(x|z)$ is postulated, with the properties $g(x|z) \geq f(x)$, for all x and z in X, and $g(x|x) = f(x)$. We then minimize $g(x|x^{k-1})$ to get x^k. Defining

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

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

Every MM algorithm, and therefore every PMA, can be viewed as an application of alternating minimization: define $\Phi(x, z) \doteq g(x|z)$. Minimizing $g(x|x^{k-1})$ to get x^k is equivalent to minimizing $\Phi(x, x^{k-1})$, while minimizing $g(x^k|z)$ is equivalent to minimizing $\Phi(x^k, z)$ and yields $z = x^k$.

3 Alternating Minimization (AM)

In this section we review the basics of alternating minimization (AM).

3.1 The AM Method

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

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

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

The alternating minimization method [9] is to minimize $\Phi(x, y^{k-1})$ to get x^{k-1} and then to minimize $\Phi(x^{k-1}, y)$ to get y^k. Clearly, the sequence $\{\Phi(x^{k-1}, y^k)\}$ is decreasing and converges to some $\beta^* \geq -\infty$. We want $\beta^* = \Phi(\hat{x}, \hat{y})$, or, at least, for $\beta^* = \beta$, where $\beta = \inf_{x,y} \Phi(x, y)$.

It is helpful to reformulate AM as a method for minimizing a function $f(x)$ of the single variable $x \in X$. For each $x \in X$, let $y(x) \in Y$ be such that $\Phi(x, y) \geq \Phi(x, y(x))$, for all $y \in Y$. Then minimizing $\Phi(x, y)$ over all $x \in X$ and $y \in Y$ is equivalent to minimizing $f(x) \doteq \Phi(x, y(x))$ over all $x \in X$. Note that $\Phi(x^{k-1}, y^k) = f(x^{k-1})$. Then the sequence $\{f(x^k)\}$ is decreasing to β^*.

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

$$d(x, x') \doteq \Phi(x, y(x')) - \Phi(x, y(x)). \tag{3.1}$$

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

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

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

3.2 The Three-Point Property

The *three-point property* (3PP) in [9] is the following: for all $x \in X$ and $y \in Y$ and for all k we have

$$\Phi(x, y^k) - \Phi(x^k, y^k) \geq d(x, x^k). \tag{3.2}$$

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

3.3 The Weak Three-Point Property

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

$$\Phi(x, y^k) - \Phi(x^k, y^{k+1}) \geq d(x, x^k). \tag{3.3}$$

3.4 Consequences of the w3PP

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

$$d(x, x^{k-1}) - d(x, x^k) \geq \Phi(x^k, y^{k+1}) - \Phi(x, y(x)). \tag{3.4}$$

Since

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

we conclude that, whenever the w3PP holds, we have

$$d(x, x^{k-1}) - d(x, x^k) \geq f(x^k) - f(x), \tag{3.5}$$

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

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

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

$$g(x|x^{k-1}) - g(x|x^k) \geq f(x^k) - f(x) \tag{3.6}$$

for all x. In the language of PMA it becomes

$$d(x, x^{k-1}) - d(x, x^k) \geq f(x^k) - f(x) \tag{3.7}$$

for all x.

4 PMA with Bregman Distances (PMAB)

Let $f : \mathbb{R}^J \to \mathbb{R}$ and $h : \mathbb{R}^J \to \mathbb{R}$ be convex and differentiable. Let $D_h(x, z)$ be the Bregman distance associated with h. At the kth step of a proximal minimization algorithm with Bregman distance (PMAB) we minimize

$$G_k(x) = f(x) + D_h(x, x^{k-1}) \tag{4.8}$$

to get x^k. It was shown in [4] that such algorithms are in the SUMMA class.

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

$$0 = \nabla f(x) + \nabla h(x) - \nabla h(x^{k-1}) \tag{4.9}$$

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

$$0 = \nabla g(x) - \nabla g(x^{k-1}) + \nabla f(x^{k-1}) \tag{4.10}$$

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

5 The Landweber Algorithm

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

$$D_f(x, y) = \|Ax - Ay\|^2, \tag{5.11}$$

so that

$$D_h(x, y) = \frac{1}{\gamma}\|x - y\|^2 - \|Ax - Ay\|^2. \tag{5.12}$$

At the kth step we differentiate

$$\|Ax - b\|^2 + \frac{1}{\gamma}\|x - x^{k-1}\|^2 - \|Ax - Ax^{k-1}\|^2, \tag{5.13}$$

to obtain

$$0 = A^T(Ax - b) + \frac{1}{\gamma}(x - x^{k-1}) - A^T(Ax - Ax^{k-1}) \tag{5.14}$$

so that

$$x^k = x^{k-1} - \gamma A^T(Ax^{k-1} - b). \tag{5.15}$$

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

References

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

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

3. 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.

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

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

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

7. Byrne, C. (2015) "The EM algorithm and related methods for iterative optimization." unpublished notes.

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

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

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

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