Nonconvex Optimization via MM Algorithms: Convergence Theory

By *Kenneth Lange[1], Joong-Ho Won[2], Alfonso Landeros[1], and Hua Zhou[1]*

Abstract: The majorization–minimization (MM) principle is an extremely general framework for deriving optimization algorithms. It includes the expectation–maximization (EM) algorithm, proximal gradient algorithm, concave–convex procedure, quadratic lower bound algorithm, and proximal distance algorithm as special cases. Besides numerous applications in statistics, optimization, and imaging, the MM principle finds wide applications in large-scale machine learning problems such as matrix completion, discriminant analysis, and nonnegative matrix factorizations. When applied to nonconvex optimization problems, MM algorithms enjoy the advantages of convexifying the objective function, separating variables, numerical stability, and ease of implementation. However, compared to the large body of literature on other optimization algorithms, the convergence analysis of MM algorithms is scattered and problem specific. This survey presents a unified treatment of the convergence of MM algorithms. With modern applications in mind, the results encompass nonsmooth objective functions and nonasymptotic analysis.

1 Background

The majorization–minimization (MM) principle for constructing optimization algorithms[1–3] finds broad range of applications in

- statistics: multidimensional scaling[4], quantile regression[5], ranking sports teams[6], variable selection[7–10], multivariate distributions[11, 12], variance component models[13], robust covariance estimation[14], and survival models[15, 16];
- optimization: geometric and sigmoid programming[17] and proximal distance algorithm[18–20];
- imaging: transmission and positron tomography[21], wavelets[22], magnetic resonance imaging, and sparse deconvolution; and
- machine learning: nonnegative matrix factorization[23], matrix completion[24, 25], clustering[26, 27], discriminant analysis[28], and support vector machines[29].

[1]University of California, Los Angeles, CA, USA
[2]Seoul National University, Seoul, South Korea

The recent book[30] and survey papers[31, 32] give a comprehensive overview of MM algorithms.

The MM principle involves majorizing the objective function $f(\boldsymbol{x})$ by a surrogate function $g(\boldsymbol{x} \mid \boldsymbol{x}_n)$ around the current iterate $\boldsymbol{x}_n$ of a search. Majorization is defined by the two conditions

$$f(\boldsymbol{x}_n) = g(\boldsymbol{x}_n \mid \boldsymbol{x}_n) \tag{1}$$

$$f(\boldsymbol{x}) \le g(\boldsymbol{x} \mid \boldsymbol{x}_n), \quad \boldsymbol{x} \neq \boldsymbol{x}_n \tag{2}$$

In other words, the surface $\boldsymbol{x} \mapsto g(\boldsymbol{x} \mid \boldsymbol{x}_n)$ lies above the surface $\boldsymbol{x} \mapsto f(\boldsymbol{x})$ and is tangent to it at the point $\boldsymbol{x} = \boldsymbol{x}_n$. Construction of the majorizing function $g(\boldsymbol{x} \mid \boldsymbol{x}_n)$ constitutes the first M of the MM algorithm. The second M of the algorithm minimizes the surrogate $g(\boldsymbol{x} \mid \boldsymbol{x}_n)$ rather than $f(\boldsymbol{x})$. If $\boldsymbol{x}_{n+1}$ denotes the minimizer of $g(\boldsymbol{x} \mid \boldsymbol{x}_n)$, then this action forces the descent property $f(\boldsymbol{x}_{n+1}) \le f(\boldsymbol{x}_n)$. This fact follows from the inequalities

$$f(\boldsymbol{x}_{n+1}) \le g(\boldsymbol{x}_{n+1} \mid \boldsymbol{x}_n) \le g(\boldsymbol{x}_n \mid \boldsymbol{x}_n) = f(\boldsymbol{x}_n)$$

reflecting the definition of $\boldsymbol{x}_{n+1}$ and the tangency condition.

The same principle applied to the maximization problems leads to the minorization–maximization algorithms that monotonically increase the objective values. The celebrated expectation–maximization (EM) algorithm in statistics is a special case of the minorization–maximization algorithm as the E-step constructs a Q-function that satisfies the minorization properties. Derivation of EM algorithm hinges upon the notion of missing data and conditional expectation while that of MM algorithm upon clever use of inequalities. For most problems where an EM algorithm exists, the MM derivation often leads to the same algorithm. Notable exceptions include the maximum-likelihood estimation (MLE) of the Dirichlet-multinomial model[11, 33] and the variance components model[13]. However, the MM principle has much wider applications as it applies to both minimization and maximization problems and does not rely on the notion of missing data.

2 Convergence Theorems

Throughout, we denote by $\mathcal{X} \subset \mathbb{R}^d$ the subset underlying our problems. All of the functions we consider have domain $\mathcal{X}$ and are extended real valued with range $\mathbb{R} \cup \{\infty\}$. The interior of set S is denoted by $\mathbf{int} \, S$, and its closure by $\mathbf{cl} \, S$.

The following concepts are useful.

Definition 1. (Effective domain). *The effective domain of a function f is defined and denoted by*

$$\mathrm{dom} f = \{\boldsymbol{x} \in \mathcal{X} : f(\boldsymbol{x}) < \infty\}$$

Definition 2. (Properness). *Function $f(\boldsymbol{x})$ is called* proper *if $\mathrm{dom} f \neq \emptyset$.*

Definition 3. (Directional derivatives). *The directional derivative of function f at $\boldsymbol{x} \in \mathcal{X}$ is defined and denoted as*

$$d_{\boldsymbol{v}} f(\boldsymbol{x}) = \lim_{t \downarrow 0} \frac{f(\boldsymbol{x} + t\boldsymbol{v}) - f(\boldsymbol{x})}{t}$$

if the limit exists.

If f is differentiable at $\boldsymbol{x}$, then $d_{\boldsymbol{v}} f(\boldsymbol{x}) = \langle \nabla f(\boldsymbol{x}), \boldsymbol{v} \rangle$.

Definition 4. (*L*-smoothness). *Function f is said to be L-smooth with respect to a norm $\| \cdot \|$ if it is differentiable on* **int** $\mathrm{dom}\, f$ *and the gradient ∇f is Lipschitz continuous with a Lipschitz constant L:*

$$\| \nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}) \| \le L \| \boldsymbol{x} - \boldsymbol{y} \|, \quad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbf{int}\, \mathrm{dom}\, f$$

It can be shown that $f(\boldsymbol{x})$ is L-smooth if and only if

$$f(\boldsymbol{x}) \le f(\boldsymbol{y}) + \langle \nabla f(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle + \frac{L}{2} \| \boldsymbol{x} - \boldsymbol{y} \|^2, \quad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbf{int}\, \mathrm{dom}\, f$$

Definition 5. (Strong convexity). *Function f is called μ-strongly convex with respect to a norm $\| \cdot \|$, $\mu \ge 0$, if $f(\boldsymbol{x}) - \frac{\mu}{2} \|\boldsymbol{x}\|^2$ is convex.*

It can be shown that if $f(\boldsymbol{x})$ is μ-strongly convex and has its minimum at $\boldsymbol{y}$, then

$$f(\boldsymbol{x}) - f(\boldsymbol{y}) \ge \frac{\mu}{2} \| \boldsymbol{x} - \boldsymbol{y} \|^2$$

Definition 6. (Tangent vector, tangent cone). *For a closed nonempty set $C \subset \mathcal{X}$, the* tangent cone *of C at $\boldsymbol{x}$ is*

$$T_C(\boldsymbol{x}) = \left\{ \boldsymbol{v} \in \mathcal{X} : \exists \{\boldsymbol{x}_n\} \subset C, \{t_n\} \subset \mathbb{R} \text{ such that } t_n \downarrow 0, \boldsymbol{x}_n \to \boldsymbol{x} \text{ and } \frac{\boldsymbol{x}_n - \boldsymbol{x}}{t_n} \to \boldsymbol{v} \right\}$$

where the notation $t_n \downarrow 0$ means that t_n approaches 0 from above. A vector $\boldsymbol{v} \in T_C(\boldsymbol{x})$ is said to be a tangent vector *of C at $\boldsymbol{x}$.*

2.1 Classical Convergence Theorem

Consider the problem of minimizing the objective function f over a closed nonempty set $C \subset \mathcal{X}$. The following is immediate from the decent property of the MM algorithms:

Proposition 1. *Let $\{\boldsymbol{x}_n\} \subset \mathcal{X}$ be the iterates generated by an MM algorithm. Assume (a) $\boldsymbol{x}_n \in C$ for each n. Then, the sequence of objective values $\{f(\boldsymbol{x}_n)\}$ monotonically decreases. Furthermore, if (b) $p^* = \inf_{\boldsymbol{x} \in C} f(\boldsymbol{x}) > -\infty$, then $\{f(\boldsymbol{x}_n)\}$ converges.*

Whether the limit is the desired minimum and whether the iterate $\{\boldsymbol{x}_n\}$ will converge to a minimizer is more subtle. For the latter, a classical theory of convergence in nonlinear optimization algorithms is due to Zangwill and Mond[34]. We first recap Zangwill's theory following the modern treatment of Luenberger and Ye[35]. Note that most of the iterative optimization algorithms, including the MM algorithms, generate a sequence $\{\boldsymbol{x}_n\}$ by mapping $\boldsymbol{x}_n \in \mathcal{X}$ to another point $\boldsymbol{x}_{n+1} \in \mathcal{X}$. For example, in MM algorithms, $\boldsymbol{x}_{n+1}$ is a point that minimizes the surrogate function $g(\boldsymbol{x}|\boldsymbol{x}_n)$ in $\mathcal{X}$. However, such a minimizer may not be unique unless the $g(\boldsymbol{x}|\boldsymbol{x}_n)$ satisfies certain assumptions. Rather, $\boldsymbol{x}_{n+1}$ is *one of the minimizers* of $g(\boldsymbol{x}|\boldsymbol{x}_n)$ and can be written as $\boldsymbol{x}_{n+1} \in \mathrm{argmin}_{\boldsymbol{x} \in C} g(\boldsymbol{x}|\boldsymbol{x}_n)$. Thus, we may in general define an algorithm map as a *set-valued map*:

Definition 7. (Algorithm map). *An algorithm map M is a mapping defined on $\mathcal{X}$ that assigns to every point $\boldsymbol{x} \in \mathcal{X}$ a subset of $\mathcal{X}$.*

Among which point $M(\boldsymbol{x}_n)$ to choose as $\boldsymbol{x}_{n+1}$ depends on the specific details of the actual optimization algorithm. If M is a single-valued map, that is, $M(\boldsymbol{x})$ is singleton for all $\boldsymbol{x} \in \mathcal{X}$, we write $\boldsymbol{x}_{n+1} = M(\boldsymbol{x}_n)$.

A desirable property of an algorithm map is closure, which extends continuity of single-valued maps to set-valued ones:

Definition 8. (Closure). *A set-valued map M from $\mathcal{X}$ to $\mathcal{Y}$ is said to be* closed *at $x \in \mathcal{X}$ if $y \in M(x)$ whenever $\{x_n\} \subset \mathcal{X}$ converges to x and $\{y_n : y_n \in M(x_n)\}$ converges to y. The map M is said to be closed on $\mathcal{X}$ if it is closed at each point of $\mathcal{X}$.*

The celebrated Zangwill's global convergence theorem is phrased in terms of an algorithm map M, a solution set Γ, and a *descent function u*:

Lemma 1. (Convergence Theorem A, Zangwill and Mond[34]). *Let the point-to-set map $M : \mathcal{X} \to \mathcal{X}$ determine an algorithm that given a point $x_0 \in \mathcal{X}$ generates the sequence $\{x_n\}$. Also, let a solution set $\Gamma \subset \mathcal{X}$ be given. Suppose that:*

1. *all points x_n are in a compact set $C \subset \mathcal{X}$;*
2. *there is a continuous function $u : \mathcal{X} \to \mathbb{R}$ such that (a) if $x \notin \Gamma$, $u(y) < u(x)$ for all $y \in M(x)$, and (b) if $x \in \Gamma$, then either the algorithm terminates or $u(y) \le u(x)$ for all $y \in M(x)$;*
3. *the map M is closed at x if $x \notin \Gamma$.*

Then, either the algorithm stops at a solution, or the limit of any convergent subsequence is a solution.

In applying Lemma 1 to specific algorithms, one usually needs to show the closure of the algorithm map M and carefully choose the solution set Γ and the descent function u. For example, in an MM algorithm, we can choose u as the objective function f and the solution set

$$\Gamma = \{x \in \mathcal{X} : f(y) \ge f(x), \; \forall y \in M(x)\}$$

for $M(x) = \operatorname{argmin}_{z \in \mathcal{X}} g(z|x)$. Since $f(y) \le f(x)$ for all $y \in M(x)$ by the descent property of MM, in fact

$$\Gamma = \{x \in \mathcal{X} : f(y) = f(x), \; \forall y \in M(x)\} =: \mathcal{P}$$

which we call a set of *no-progress points*. The final requirement that $\{x_n\}$ is contained within a compact set is satisfied whenever f is lower semicontinuous and coercive.

We summarize the above discussion as the following proposition: see also Proposition 8 of Keys *et al.*[20].

Proposition 2. (Global convergence to no-progress points). *Suppose that the objective f is lower semicontinuous and coercive, and the algorithm map M defined by the MM algorithm is closed. Then, all the limit points of the iterates $x_{n+1} \in M(x_n)$ generated by the MM algorithm are no-progress points.*

This general result is slightly disappointing. Even though the objective values do not change within $\mathcal{P}$, the iterate $\{x_n\}$ may not even converge – it may cycle through distinct no-progress points.

Example 1. (EM algorithm). As a classical example of cycling, Vaida[36] showed that in minimizing

$$f(\rho, \sigma^2) = 8 \log \sigma^2 + \frac{18}{\sigma^2} + 2 \log(1 - \log \rho^2) + \frac{4}{\sigma^2(1 - \rho^2)}$$

over $\sigma^2 \le 0$ and $-1 \le \rho \le 1$ (this objective function originates from the maximum-likelihood estimation of the variance and correlation coefficient of bivariate normal data with missing observations), the following particular surrogate function

$$g(\rho, \sigma^2 \mid \rho_n, \sigma_n^2) = f(\rho, \sigma^2) + 2 \left(\log \frac{\sigma^2(1 - \rho^2)}{\sigma_n^2(1 - \rho_n^2)} + \frac{\sigma_n^2(1 - \rho_n^2)}{\sigma^2(1 - \rho^2)} - 1 \right)$$

obtained by applying the EM algorithm, a special case of the MM algorithms, has two symmetric minima, $(\sigma_{n+1}^2, \rho_{n+1}) = (3, \pm\sqrt{2/3 - \sigma_n^2(1 - \rho_n^2)/6})$. If we take $\sigma_0^2 = 3$ and

$$\rho_{n+1} = -\text{sgn}(\rho_n)\sqrt{2/3 - 3(1 - \rho_n^2)/6}$$

then the sequence $\{(\sigma_n^2, \rho_n)\}$ oscillates between two minima $(3, \pm1/\sqrt{3})$ of f in the limit.

Although the above cycling can be considered desirable as it reveals multiple optima, the next example shows that this is not always the case:

Example 2. (Generalized CCA). The popular MAXDIFF criterion[37−39] for generalizing the CCA into $m > 2$ sets of (partially) orthogonal matrices solves

$$\text{maximize} \sum_{i<j} \text{tr}(O_i^T A_i^T A_j O_j) \text{ subject to } O_i^T O_i = I_r, \; i = 1, \dots, m \tag{3}$$

where I_r is an $r \times r$ identity matrix, and $O_i \in \mathbb{R}^{d_i \times r}$; $A_i \in \mathbb{R}^{n \times d_i}$ are n observations of variables of possibly different dimensions. A standard algorithm for solving the MAXDIFF problem is Ten Berge's block relaxation algorithm[38, 40], shown as Algorithm 1. This is an MM algorithm (here minorization−maximization), since at the update of the ith block in the kth sweep, the surrogate function

$$g(O_1, \dots, O_m | O_1^{k+1}, \dots, O_{i-1}^{k+1}, O_i^k, \dots, O_m^k)$$
$$= \frac{1}{2} \sum_{i=1}^m \text{tr} \left[O_i^T \left(\sum_{j=1}^{i-1} A_i^T A_j O_j^{k+1} + \sum_{j=i+1}^m A_i^T A_j O_j^k \right) \right]$$

minorizes the objective function of problem (3) at $(O_1^{k+1}, \dots, O_{i-1}^{k+1}, O_i^k, \dots, O_m^k)$ and is maximized based on the von Neuman−Fan inequality

$$\text{tr}(A^T B) \le \sum_l \sigma_l(A)\sigma_l(B)$$

which holds for any two matrices A and B of the same dimensions with the lth largest singular values $\sigma_l(A)$ and $\sigma_l(B)$, respectively; equality is attained when A and B share a simultaneous ordered Singular Value Decomposition (SVD) [41].

While each iteration monotonically improves the objective function. Won *et al.*[42] show that Algorithm 1 may oscillate between suboptimal no-progress points. Set $m = 3$, $d_1 = d_2 = d_3 = d = r$ and $A_1 = [I_d, I_d, 0]^T$, $A_2 = [-I_d, 0, I_d]^T$, and $A_3 = [0, I_d, I_d]^T$[43]. If Algorithm 1 is initialized with (J, K, J) where

$$J = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \quad \text{and} \quad K = \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 0 & 0 \end{bmatrix},$$

then both $J - K$ and $J + K$ have rank 1, and we see that $-K$ is one of the maximizers of $\text{tr}[O^T(J - K)]$, and likewise, J maximizes $\text{tr}[O^T(J + K)]$. Taking these values as the outputs of Line 5 of Algorithm 1, we have the following cycling sequence at the end of each sweep:

$$(J, K, J) \to (-K, J, -K) \to (-J, -K, -J) \to (K, -J, K) \to (J, K, J) \to \cdots$$

All four limit points yield the same objective values of 1. However, the global maximum of f can be shown to be 3.

The main reason for this oscillatory behavior is that the map $\boldsymbol{B} = \sum_{j \neq i} \boldsymbol{A}_i^T \boldsymbol{A}_j \boldsymbol{O}_j \mapsto \boldsymbol{P}_i \boldsymbol{Q}_i^T$ in Lines 5 and 6 is set valued. If $\boldsymbol{B}$ is rank deficient, any orthonormal basis of the null space of $\boldsymbol{B}^T$ (*resp.* $\boldsymbol{B}$) can be chosen as left (*resp.* right) singular vectors corresponding to the zero singular value. Furthermore, the product $\boldsymbol{P}_i \boldsymbol{Q}_i^T$ may not be unique [44, Proposition 7].

More satisfying "solution sets" are in order.

- Fixed points:
$$\mathcal{F} = \{\boldsymbol{x} \in \mathcal{X} \, : \, \boldsymbol{x} = M(\boldsymbol{x})\}$$

  if M is single valued.
- Stationary points:
$$S = \{\boldsymbol{x} \in \mathcal{X} \, : \, d_{\boldsymbol{v}} f(\boldsymbol{x}) \geq 0, \text{ for all tangent vectors } \boldsymbol{v} \text{ of } C \text{ at } \boldsymbol{x}\}$$

Algorithm 1 Ten Berge's algorithm for generalized CCA

1: Initialize $\boldsymbol{O}_1, \dots, \boldsymbol{O}_m$
2: For $k = 1, 2, \dots$
3:     For $i = 1, \dots, m$
4:         Set $\boldsymbol{B} = \sum_{j \neq i} \boldsymbol{A}_i^T \boldsymbol{A}_j \boldsymbol{O}_j$
5:         Compute SVD of $\boldsymbol{B}$ as $\boldsymbol{P}_i \boldsymbol{D}_i \boldsymbol{Q}_i^T$
6:         Set $\boldsymbol{O}_i = \boldsymbol{P}_i \boldsymbol{Q}_i^T$
7:     End For
8:     If there is no progress, then break
9: End For
10: Return $(\boldsymbol{O}_1, \dots, \boldsymbol{O}_m)$

All fixed points are no-progress points, that is, $\mathcal{F} \subset \mathcal{P}$, but not vice versa. Note that $\boldsymbol{y} \in S$ is a *necessary* condition that $\boldsymbol{y}$ is a local minimizer of f in C. No-progress points and fixed points depend on the algorithm map M, whereas the stationary points depend on the problem itself. To make M single valued, note that any convex (and weakly convex) surrogate $g(\boldsymbol{x}|\boldsymbol{x}_n)$ can be made strongly convex, thus attains a unique minimum, by adding the viscosity penalty $\frac{\mu}{2} \| \boldsymbol{x} - \boldsymbol{x}_n \|^2$ majorizing 0; see Section 2.4. If M is closed and single valued, then it is continuous.

The classical global convergence results for MM algorithms[30, 45], which we summarize below, hinge on continuity of the map M:

Proposition 3. *If the MM algorithm map M is continuous, then $\mathcal{F} = \mathcal{P}$ and $\mathcal{F}$ is closed.*

Proposition 4. *If (i) f is continuous, (ii) f is coercive, or the set $\{\boldsymbol{x} : f(\boldsymbol{x}) \leq f(\boldsymbol{x}_0)\}$ is compact, and (iii) the algorithm map M is continuous, then every limit point of an MM sequence $\{\boldsymbol{x}_n\}$ is a fixed point of M. Furthermore, $\lim_{n \to \infty} \text{dist}(\boldsymbol{x}_n, \mathcal{F}) = 0$.*

Proposition 5. *Under the same assumptions of Proposition 4, the MM sequence $\{\boldsymbol{x}_n\}$ satisfies*
$$\lim_{n \to \infty} \| \boldsymbol{x}_{n+1} - \boldsymbol{x}_n \| = 0$$

Furthermore, the set $\mathcal{T}$ of the limit points of $\{\boldsymbol{x}_n\}$ is compact and connected.

Note Proposition 4 states that $\mathcal{T} \subset \mathcal{F}$. Proposition 5 ensures there is no cycling.

Connecting the fixed points $\mathcal{F}$, which coincides with the no-progress points $\mathcal{P}$ for continuous M, with the stationary points $\mathcal{S}$ needs more assumptions. To equate stationary points of f to those of $g(\cdot|\boldsymbol{x})$, we require a stronger tangency condition than the usual tangency condition (1):

Definition 9. (Strong tangency). *An MM surrogate function $g(\cdot|\cdot)$ is said to be* strongly tangent *to f at $\boldsymbol{x} \in C$ if $d_v g(\boldsymbol{x}|\boldsymbol{x}) = d_v f(\boldsymbol{x})$ for all $\boldsymbol{x} \in C$ and tangent vector $\boldsymbol{v}$ in C at $\boldsymbol{x}$.*

Proposition 6. *Suppose that (i) the surrogate function $g(\boldsymbol{y}|\boldsymbol{x})$ is strongly tangent to the objective function f, (ii) the algorithm map M is closed and single valued, and (iii) stationary points and minimizers of $g(\boldsymbol{y}|\boldsymbol{x})$ are equivalent. Then, $\mathcal{F} = \mathcal{P} = \mathcal{S}$, that is, the sets of fixed points, no-progress points, and stationary points of f coincide.*

Proposition 7. *In addition to the assumptions of Proposition 6, if $\mathcal{S}$, the set of all stationary points of the objective function f, consists of isolated points, then the set $\mathcal{T}$ of the limit points of the MM sequence $\{\boldsymbol{x}_n\}$ is singleton, that is, an MM sequence $\{\boldsymbol{x}_n\}$ possesses a limit, and that limit is a stationary point of f as well as a fixed point of M.*

Strong tangency holds when $g(\boldsymbol{y}|\boldsymbol{x}) = f(\boldsymbol{x}) + h(\boldsymbol{y}|\boldsymbol{x})$ and $h(\boldsymbol{y}|\boldsymbol{x})$ is differentiable with $\nabla h(\boldsymbol{x}|\boldsymbol{x}) = \boldsymbol{0}$. See Vaida[36] and Yu *et al.*[46] for examples of these results in action.

We next present results that extend to nonasymptotic analysis and more general settings such as nonsmooth objectives.

2.2 Smooth Objective Functions

The following proposition gives a weak form of convergence for MM algorithms. The proposition features minimization and majorization by Lipschitz smooth functions.

Proposition 8. *Let $f(\boldsymbol{x})$ be a coercive differentiable function majorized by a uniformly L-Lipschitz surrogate $g(\boldsymbol{x} \mid \boldsymbol{x}_n)$ anchored at $\boldsymbol{x}_n$. If $\boldsymbol{y}$ denotes a minimum point of $f(\boldsymbol{x})$, then the iterates $\boldsymbol{x}_n$ delivered by the corresponding MM algorithm satisfy the sublinear bound*

$$\min_{0 \leq k \leq n} \| \nabla f(\boldsymbol{x}_k)\|^2 \leq \frac{2L}{n+1}[f(\boldsymbol{x}_0) - f(\boldsymbol{y})] \tag{4}$$

When $f(\boldsymbol{x})$ is continuously differentiable, any limit point of the sequence $\boldsymbol{x}_n$ is a stationary point of $f(\boldsymbol{x})$.

Proof. Given that the surrogate $g(\boldsymbol{x} \mid \boldsymbol{x}_n)$ satisfies the tangency condition $\nabla g(\boldsymbol{x}_n \mid \boldsymbol{x}_n) = \nabla f(\boldsymbol{x}_n)$, the L-smoothness assumption entails the quadratic upper bound

$$\begin{aligned}
f(\boldsymbol{x}_{n+1}) - f(\boldsymbol{x}_n) &\leq g(\boldsymbol{x}_{n+1} \mid \boldsymbol{x}_n) - g(\boldsymbol{x}_n \mid \boldsymbol{x}_n) \\
&\leq g(\boldsymbol{x} \mid \boldsymbol{x}_n) - g(\boldsymbol{x}_n \mid \boldsymbol{x}_n) \\
&\leq \langle \nabla g(\boldsymbol{x}_n \mid \boldsymbol{x}_n), \boldsymbol{x} - \boldsymbol{x}_n \rangle + \frac{L}{2} \| \boldsymbol{x} - \boldsymbol{x}_n\|^2 \\
&= \langle \nabla f(\boldsymbol{x}_n), \boldsymbol{x} - \boldsymbol{x}_n \rangle + \frac{L}{2} \| \boldsymbol{x} - \boldsymbol{x}_n\|^2
\end{aligned}$$

for any $\boldsymbol{x}$. The choice $\boldsymbol{x} = \boldsymbol{x}_n - L^{-1}\nabla f(\boldsymbol{x}_n)$ yields the sufficient decrease condition

$$f(\boldsymbol{x}_n) - f(\boldsymbol{x}_{n+1}) \geq \frac{1}{2L} \| \nabla f(\boldsymbol{x}_n)\|^2 \tag{5}$$

A simple telescoping argument now gives

$$\frac{n+1}{2L} \min_{0 \le k \le n} \| \nabla f(\boldsymbol{x}_k)\|^2 \le \frac{1}{2L} \sum_{k=0}^{n} \| \nabla f(\boldsymbol{x}_k)\|^2$$
$$\le f(\boldsymbol{x}_0) - f(\boldsymbol{x}_{n+1})$$
$$\le f(\boldsymbol{x}_0) - f(\boldsymbol{y})$$

which is equivalent to the bound (4). The second assertion follows directly from condition (5), the convergence of the sequence $f(\boldsymbol{x}_n)$, and the continuity of $\nabla f(\boldsymbol{x})$.

As a prelude to our next result, we state and prove a simple result of independent interest.

Proposition 9. *Suppose that $f(\boldsymbol{x})$ is convex with surrogate $g(\boldsymbol{z} \mid \boldsymbol{x})$ at the point $\boldsymbol{x}$. Then, $f(\boldsymbol{x})$ is differentiable at $\boldsymbol{x}$, and $\nabla f(\boldsymbol{x})$ equals $\nabla g(\boldsymbol{x} \mid \boldsymbol{x})$ wherever $\nabla g(\boldsymbol{x} \mid \boldsymbol{x})$ exists.*

Proof. Suppose that $\boldsymbol{x}$ is such a point. Let $\boldsymbol{v} \in \partial f(\boldsymbol{x})$ be a subgradient of $f(\boldsymbol{x})$. It suffices to show that $\boldsymbol{v}$ is uniquely determined as $\boldsymbol{v} = \nabla g(\boldsymbol{x} \mid \boldsymbol{x})$. For any direction $\boldsymbol{u}$, consider the forward difference quotient

$$\frac{g(\boldsymbol{x} + t\boldsymbol{u} \mid \boldsymbol{x}) - g(\boldsymbol{x} \mid \boldsymbol{x})}{t} \ge \frac{f(\boldsymbol{x} + t\boldsymbol{u}) - f(\boldsymbol{x})}{t} \ge \langle \boldsymbol{v}, \boldsymbol{u} \rangle$$

Taking limits produces $\langle \nabla g(\boldsymbol{x} \mid \boldsymbol{x}), \boldsymbol{u} \rangle \ge \langle \boldsymbol{v}, \boldsymbol{u} \rangle$. This cannot be true for all $\boldsymbol{u}$ unless the condition $\boldsymbol{v} = \nabla g(\boldsymbol{x} \mid \boldsymbol{x})$ holds.

Imposing strong convexity on $f(\boldsymbol{x})$ recovers linear convergence. The ratio $\frac{\mu}{L}$ makes an appearance, but, unlike convergence theorems for gradient descent, this ratio is not the condition number of either the objective or the surrogate.

Proposition 10. *Let $f(\boldsymbol{x})$ be a μ-strongly convex function majorized by a uniformly L-Lipschitz surrogate $g(\boldsymbol{x} \mid \boldsymbol{x}_n)$. If the global minimum occurs at $\boldsymbol{y}$, then the MM iterates $\boldsymbol{x}_n$ satisfy*

$$f(\boldsymbol{x}_n) - f(\boldsymbol{y}) \le \left[1 - \left(\frac{\mu}{L}\right)^2\right]^n [f(\boldsymbol{x}_0) - f(\boldsymbol{y})]$$

thus establishing linear convergence of $\boldsymbol{x}_n$ to $\boldsymbol{y}$.

Proof. Existence and uniqueness of $\boldsymbol{y}$ follow from strong convexity. Because $\nabla g(\boldsymbol{y} \mid \boldsymbol{y}) = \boldsymbol{0}$, the smoothness of $g(\boldsymbol{x} \mid \boldsymbol{y})$ gives the quadratic upper bound

$$f(\boldsymbol{x}) - f(\boldsymbol{y}) \le g(\boldsymbol{x} \mid \boldsymbol{y}) - g(\boldsymbol{y} \mid \boldsymbol{y})$$
$$\le \langle \nabla g(\boldsymbol{y} \mid \boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle + \frac{L}{2} \| \boldsymbol{x} - \boldsymbol{y}\|^2 \qquad (6)$$
$$= \frac{L}{2} \| \boldsymbol{x} - \boldsymbol{y}\|^2$$

which incidentally implies $\mu \le L$. By the previous proposition, $f(\boldsymbol{x})$ is everywhere differentiable with $\nabla f(\boldsymbol{x}) = \nabla g(\boldsymbol{x} \mid \boldsymbol{x})$. In view of the strong convexity assumption, we have the lower bound

$$\| \nabla f(\boldsymbol{x}) \| \cdot \| \boldsymbol{y} - \boldsymbol{x} \| \ge -\langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle$$
$$\ge f(\boldsymbol{y}) - f(\boldsymbol{x}) - \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle \qquad (7)$$
$$\ge \frac{\mu}{2} \| \boldsymbol{y} - \boldsymbol{x}\|^2$$

It follows that $\| \nabla f(\boldsymbol{x}) \| \geq \frac{\mu}{2} \| \boldsymbol{y} - \boldsymbol{x} \|$. Combining inequalities (6) and (7) furnishes the Polyak–Łojasiewicz (PL) bound

$$\| \nabla f(\boldsymbol{x}) \|^2 \geq \frac{\mu^2}{2L}[f(\boldsymbol{x}) - f(\boldsymbol{y})]$$

We now turn to the MM iterates and take $\boldsymbol{x} = \boldsymbol{x}_n - \frac{1}{L}\nabla f(\boldsymbol{x}_n)$. The PL inequality implies

$$
\begin{aligned}
f(\boldsymbol{x}_{n+1}) - f(\boldsymbol{x}_n) &\leq g(\boldsymbol{x}_{n+1} \mid \boldsymbol{x}_n) - g(\boldsymbol{x}_n \mid \boldsymbol{x}_n) \\
&\leq g(\boldsymbol{x} \mid \boldsymbol{x}_n) - g(\boldsymbol{x}_n \mid \boldsymbol{x}_n) \\
&\leq \langle \nabla g(\boldsymbol{x}_n \mid \boldsymbol{x}_n), -\frac{1}{L}\nabla g(\boldsymbol{x}_n \mid \boldsymbol{x}_n)\rangle + \frac{L}{2}\|\frac{1}{L}\nabla g(\boldsymbol{x}_n \mid \boldsymbol{x}_n) \|^2 \\
&= -\frac{1}{2L} \| \nabla f(\boldsymbol{x}_n)\|^2 \\
&\leq -\frac{2\mu^2}{L^2}[f(\boldsymbol{x}_n) - f(\boldsymbol{y})]
\end{aligned}
$$

Subtracting $f(\boldsymbol{y})$ from both sides of the previous inequality and rearranging gives

$$f(\boldsymbol{x}_{n+1}) - f(\boldsymbol{y}) \leq \left[1 - \frac{\mu^2}{2L^2}\right][f(\boldsymbol{x}_n) - f(\boldsymbol{y})]$$

Iteration of this inequality yields the claimed linear convergence.

2.3 Nonsmooth Objective Functions

Consider an MM minimization algorithm with objective $f(\boldsymbol{x})$ and surrogates $g(\boldsymbol{x} \mid \boldsymbol{x}_n)$. If $f(\boldsymbol{x})$ is coercive and continuous, and the $g(\boldsymbol{x} \mid \boldsymbol{x}_n)$ are μ-strongly convex, then we know that the MM iterates $\boldsymbol{x}_{n+1} = \operatorname{argmin}_{\boldsymbol{x}} g(\boldsymbol{x} \mid \boldsymbol{x}_n)$ remain within the compact sublevel set $S = \{\boldsymbol{x} : f(\boldsymbol{x}) \leq f(\boldsymbol{x}_0)\}$[30]. Furthermore, the strong convexity inequality

$$f(\boldsymbol{x}_n) - f(\boldsymbol{x}_{n+1}) \geq g(\boldsymbol{x}_n \mid \boldsymbol{x}_n) - g(\boldsymbol{x}_{n+1} \mid \boldsymbol{x}_n) \geq \frac{\mu}{2} \| \boldsymbol{x}_n - \boldsymbol{x}_{n+1}\|^2 \tag{8}$$

implies that

$$\sum_{n=0}^{\infty} \| \boldsymbol{x}_n - \boldsymbol{x}_{n+1}\|^2 \leq \frac{1}{\mu}[f(\boldsymbol{x}_0) - \bar{f}]$$

where $\bar{f} = \lim_{n\to\infty} f(\boldsymbol{x}_n)$. It follows from a well-known theorem of Ostrowski[30] that the set W of limit points of the $\boldsymbol{x}_n$ is compact and connected. It is also easy to show that $f(\boldsymbol{x})$ takes the constant value $\bar{f}$ on W and that $\lim_{n\to\infty} \operatorname{dist}(\boldsymbol{x}_n, W) = 0$.

We will need the concept of a Fréchet subdifferential. If $f(\boldsymbol{x})$ is a function mapping $\mathbb{R}^p$ into $\mathbb{R} \cup \{+\infty\}$, then its Fréchet subdifferential at $\boldsymbol{x} \in \operatorname{dom} f$ is the set

$$\partial^F f(\boldsymbol{x}) = \left\{ \boldsymbol{v} : \liminf_{\boldsymbol{y}\to\boldsymbol{x}}\frac{f(\boldsymbol{y}) - f(\boldsymbol{x}) - \boldsymbol{v}^t(\boldsymbol{y} - \boldsymbol{x})}{\| \boldsymbol{y} - \boldsymbol{x} \|} \geq 0 \right\}$$

The set $\partial^F f(\boldsymbol{x})$ is closed, convex, and possibly empty. If $f(\boldsymbol{x})$ is convex, then $\partial^F f(\boldsymbol{x})$ reduces to its convex subdifferential. If $f(\boldsymbol{x})$ is differentiable, then $\partial^F f(\boldsymbol{x})$ reduces to its ordinary differential. At a local minimum $\boldsymbol{x}$, Fermat's rule $\boldsymbol{0} \in \partial^F f(\boldsymbol{x})$ holds.

Proposition 11. *In an MM algorithm, suppose that $f(x)$ is coercive, $g(x \mid x_n)$ is differentiable, and the algorithm map $M(x)$ is closed. Then, all points z of the convergence set W are critical in the sense that $0 \in \partial^F(-f)(z)$.*

Proof. Let the subsequence x_{n_m} of the MM sequence $x_{n+1} \in M(x_n)$ converge to $z \in W$. By passing a sub-subsequence if necessary, we may suppose that x_{n_m+1} converges to y. Owing to our closedness assumption, $y \in M(z)$. Given that $f(y) = f(z)$, it is obvious that z also minimizes $g(x \mid z)$ and that $0 = \nabla g(z \mid z)$. Since the difference $h(x \mid z) = g(x \mid z) - f(x)$ achieves it minimum at $x = z$, the Fréchet subdifferential $\partial^F h(x \mid z)$ satisfies

$$0 \in \partial^F h(z \mid z) = \nabla g(z \mid z) + \partial^F(-f)(z)$$

It follows that $0 \in \partial^F(-f)(z)$. ∎

We will also need to invoke Łojasiewicz's inequality. This deep result depends on some rather arcane algebraic geometry[47, 48]. It applies to semialgebraic functions and their more inclusive cousins semianalytic functions and subanalytic functions. For simplicity, we focus on semialgebraic functions. The class of semialgebraic subsets of $\mathbb{R}^p$ is the smallest class such that:

a) It contains all sets of the form $\{x : q(x) > 0\}$ for a polynomial $q(x)$ in p variables.
b) It is closed under the formation of finite unions, finite intersections, and set complementation.

A function $a : \mathbb{R}^p \mapsto \mathbb{R}^r$ is said to be semialgebraic if its graph is a semialgebraic set of $\mathbb{R}^{p+r}$. The class of real-valued semialgebraic contains all polynomials $p(x)$. It is closed under the formation of sums, products, absolute values, reciprocals when $a(x) \neq 0$, roots when $a(x) \geq 0$, and maxima $\max\{a(x), b(x)\}$ and minima $\min\{a(x), b(x)\}$. For our purposes, it is important to note that $\mathrm{dist}(x, S)$ is a semialgebraic function whenever S is a semialgebraic set.

Łojasiewicz's inequality in its modern form[49] requires that $f(x)$ be continuous and subanalytic with a closed domain. If z is a critical point of $f(x)$, then

$$|f(x) - f(z)|^{\theta(z)} \leq c(z) \parallel v \parallel$$

for some constant $c(z)$, all x in some open ball $B_{r(z)}(z)$ around z of radius $r(z)$, and all v in $\partial^F f(x)$. This inequality applies to semialgebraic functions since they are automatically subanalytic. We apply Łojasiewicz's inequality to the points in the limit set W.

2.3.1 MM convergence for semialgebraic functions

Proposition 12. *Suppose that $f(x)$ is coercive, continuous, and subanalytic and all $g(x \mid x_n)$ are continuous, μ-strongly convex, and satisfy the Lipschitz condition*

$$\parallel \nabla g(u \mid x_n) - \nabla g(v \mid x_n) \parallel \leq L \parallel u - v \parallel$$

on the compact sublevel set $\{x : f(x) \leq f(x_0)\}$. Then, the MM iterates $x_{n+1} = \mathrm{argmin}_x g(x \mid x_n)$ converge to a critical point in W.

Proof. Because $h(x \mid y) = g(x \mid y) - f(x)$ achieves it minimum at $x = y$, the Fréchet subdifferential $\partial^F h(x \mid y)$ satisfies

$$0 \in \partial^F h(y \mid y) = \nabla g(y \mid y) + \partial^F(-f)(y).$$

It follows that $-\nabla g(\boldsymbol{y} \mid \boldsymbol{y}) \in \partial^F(-f)(\boldsymbol{y})$. By assumption

$$\| \nabla g(\boldsymbol{u} \mid \boldsymbol{x}_n) - \nabla g(\boldsymbol{v} \mid \boldsymbol{x}_n) \| \le L \| \boldsymbol{u} - \boldsymbol{v} \|$$

for all $\boldsymbol{u}$ and $\boldsymbol{v}$ and $\boldsymbol{x}_n$. In particular, because $\nabla g(\boldsymbol{x}_{n+1} \mid \boldsymbol{x}_n) = \boldsymbol{0}$, we have

$$\| \nabla g(\boldsymbol{x}_n \mid \boldsymbol{x}_n) \| \le L \| \boldsymbol{x}_{n+1} - \boldsymbol{x}_n \| \tag{9}$$

According to the Łojasiewicz inequality applied for the subanalytic function $\bar{f} - f(\boldsymbol{x})$, for each $\boldsymbol{z} \in W$ there exists a radius $r(\boldsymbol{z})$ and an exponent $\theta(\boldsymbol{z}) \in [0, 1)$ with

$$|f(\boldsymbol{u}) - f(\boldsymbol{z})|^{\theta(\boldsymbol{z})} = |\bar{f} - f(\boldsymbol{u}) - \bar{f} + \bar{f}|^{\theta(\boldsymbol{z})} \le c(\boldsymbol{z}) \| \boldsymbol{v} \|$$

for all $\boldsymbol{u}$ in the open ball $B_{r(\boldsymbol{z})}(\boldsymbol{z})$ around $\boldsymbol{z}$ of radius $r(\boldsymbol{z})$ and all $\boldsymbol{v} \in \partial^F(\bar{f} - f)(\boldsymbol{u}) = \partial^F(-f)(\boldsymbol{u})$. We apply this inequality to $\boldsymbol{u} = \boldsymbol{x}_n$ and $\boldsymbol{v} = -\nabla g(\boldsymbol{x}_n \mid \boldsymbol{x}_n)$. In doing so, we would like to assume that the exponent $\theta(\boldsymbol{z})$ and constant $c(\boldsymbol{z})$ do not depend on $\boldsymbol{z}$. With this end in mind, cover W by a finite number of balls $B_{r(\boldsymbol{z}_i)}(\boldsymbol{z}_i)$ and take $\theta = \max_i \theta(\boldsymbol{z}_i) < 1$ and $c = \max_i c(\boldsymbol{z}_i)$. For a sufficiently large N, every $\boldsymbol{x}_n$ with $n \ge N$ falls within one of these balls and satisfies $|\bar{f} - f(\boldsymbol{x}_n)| < 1$. Without loss of generality assume $N = 0$. The Łojasiewicz inequality now entails

$$|\bar{f} - f(\boldsymbol{x}_n)|^{\theta} \le c \| \nabla g(\boldsymbol{x}_n \mid \boldsymbol{x}_n) \| \tag{10}$$

In combination with the concavity of the function $t^{1-\theta}$ on $[0, \infty)$, inequalities (8), (9), and (10) imply

$$[f(\boldsymbol{x}_n) - \bar{f}]^{1-\theta} - [f(\boldsymbol{x}_{n+1}) - \bar{f}]^{1-\theta} \ge \frac{1-\theta}{[f(\boldsymbol{x}_n) - \bar{f}]^{\theta}}[f(\boldsymbol{x}_n) - f(\boldsymbol{x}_{n+1})]$$

$$\ge \frac{1-\theta}{c \| \nabla g(\boldsymbol{x}_n \mid \boldsymbol{x}_n) \|} \frac{\mu}{2} \| \boldsymbol{x}_{n+1} - \boldsymbol{x}_n \|^2$$

$$\ge \frac{(1-\theta)\mu}{2cL} \| \boldsymbol{x}_{n+1} - \boldsymbol{x}_n \|$$

Rearranging this inequality and summing over n yield

$$\sum_{n=0}^{\infty} \| \boldsymbol{x}_{n+1} - \boldsymbol{x}_n \| \le \frac{2cL}{(1-\theta)\mu}[f(\boldsymbol{x}_0) - \bar{f}]^{1-\theta}$$

Thus, the sequence $\boldsymbol{x}_n$ is a fast Cauchy sequence and converges to a unique limit in W.

2.4 A Proximal Trick to Prevent Cycling

Consider minimizing a function $f(\boldsymbol{x})$ bounded below and possibly subject to constraints. The MM principle involves constructing a surrogate function $g(\boldsymbol{x} \mid \boldsymbol{x}_n)$ that majorizes $f(\boldsymbol{x})$ around $\boldsymbol{x}_n$. For any $\rho > 0$, adding the penalty $(\rho/2) \| \boldsymbol{x} - \boldsymbol{x}_n \|^2$ to the surrogate produces a new surrogate

$$g(\boldsymbol{x} \mid \boldsymbol{x}_n) + \frac{\rho}{2} \| \boldsymbol{x} - \boldsymbol{x}_n \|^2$$

Rearranging the inequality

$$g(\boldsymbol{x}_{n+1} \mid \boldsymbol{x}_n) + \frac{\rho}{2} \| \boldsymbol{x}_{n+1} - \boldsymbol{x}_n \|^2 \le g(\boldsymbol{x}_n \mid \boldsymbol{x}_n)$$

yields

$$\frac{\rho}{2} \| \boldsymbol{x}_{n+1} - \boldsymbol{x}_n \|^2 \le g(\boldsymbol{x}_n \mid \boldsymbol{x}_n) - g(\boldsymbol{x}_{n+1} \mid \boldsymbol{x}_n) \le f(\boldsymbol{x}_n) - f(\boldsymbol{x}_{n+1})$$

Thus, the MM iterates induced by the new surrogate satisfy

$$\lim_{n \to \infty} \| \boldsymbol{x}_{n+1} - \boldsymbol{x}_n \| = 0$$

This property is inconsistent with algorithm cycling between distant limit points.

3 Paracontraction

Another useful tool for proving iterate convergence of MM algorithms is paracontraction. Recall that a map $T : \mathcal{X} \to \mathbb{R}^d$ is contractive with respect to a norm $\| \boldsymbol{x} \|$ if $\| T(\boldsymbol{y}) - T(\boldsymbol{z}) \| < \| \boldsymbol{y} - \boldsymbol{z} \|$ for all $\boldsymbol{y} \neq \boldsymbol{z}$ in $\mathcal{X}$. It is strictly contractive if there exists a constant $c \in [0, 1)$ with $\| T(\boldsymbol{y}) - T(\boldsymbol{z}) \| \leq c \| \boldsymbol{y} - \boldsymbol{z} \|$ for all such pairs. If $c = 1$, then the map is nonexpansive.

Definition 10. (Paracontractive map). *A map $T : \mathcal{X} \to \mathbb{R}^d$ is said to be* paracontractive *if for every fixed point $\boldsymbol{y}$ of T (i.e., $\boldsymbol{y} = T(\boldsymbol{y})$), the inequality $\| T(\boldsymbol{x}) - \boldsymbol{y} \| < \| \boldsymbol{x} - \boldsymbol{y} \|$ holds unless $\boldsymbol{x}$ is itself a fixed point.*

A strictly contractive map is contractive, and a contractive map is paracontractive.

An important result regarding paracontractive maps is the theorem of Elsner, Koltract, and Neumann[50], which states that whenever a continuous paracontractive map T possesses one or more fixed points, then the sequence of iterates $\boldsymbol{x}_{n+1} = T(\boldsymbol{x}_n)$ converges to a fixed point regardless of the initial point $\boldsymbol{x}_0$. More formal statement is as follows:

Proposition 13. *Suppose that the continuous maps $T_0, \cdots, T_{r-1}$ of a set into itself are paracontractive under the norm $\| x \|$. Let F_i denote the set of fixed points of T_i. If the intersection $F = \cap_{i=0}^{r-1} F_i$ is nonempty, then the sequence*

$$\boldsymbol{x}_{n+1} = T_{n \bmod r}(\boldsymbol{x}_n)$$

converges to a limit in F. In particular, if $r = 1$ and $T = T_0$ has a nonempty set of fixed points F, then $\boldsymbol{x}_{n+1} = T(\boldsymbol{x}_n)$ converges to a point in F.

A simple proof is given in Lange[51].

Proposition 13 converts the task of proving convergence of MM iterates to that of showing (i) continuity, (ii) paracontractivity, and (iii) existence of a fixed point, of the MM algorithm map, and that (iv) any fixed point is a stationary point of the objective. A nice example is the recent work by Won *et al.*[52] on Euclidean projection onto the Minkowski sum of sets. The Minkowski sum of two sets A and B in $\mathbb{R}^d$ is

$$A + B = \{ \boldsymbol{a} + \boldsymbol{b} : \boldsymbol{a} \in A, \ \boldsymbol{b} \in B \}$$

It is easy to show that $A + B$ is convex whenever A and B are both convex and is closed if at least one of the two sets is compact and the other is closed. When $A + B$ is closed with A and B convex, we may employ a block descent algorithm, an instance of MM algorithms, for finding the closest point to $\boldsymbol{x} \notin A + B$, which consists of alternating

$$\begin{aligned} \boldsymbol{b}_{n+1} &= P_B(\boldsymbol{x} - \boldsymbol{a}_n) \\ \boldsymbol{a}_{n+1} &= P_A(\boldsymbol{x} - \boldsymbol{b}_{n+1}) \end{aligned} \tag{11}$$

assuming that the projection operators P_A and P_B onto A and B are both known or easy to compute.

In order to show that the sequence $\{ \boldsymbol{a}_n + \boldsymbol{b}_n \}$ converges to the closest point using Proposition 13, we first need to show the continuity of the map

$$T(\boldsymbol{a}) = P_A[\boldsymbol{x} - P_B(\boldsymbol{x} - \boldsymbol{a})]$$

The obtuse angle property of Euclidean projection [51, Example 6.5.3] yields

$$\langle a - P_A(a), P_A(\tilde{a}) - P_A(a) \rangle \leq 0$$
$$\langle \tilde{a} - P_A(\tilde{a}), P_A(a) - P_A(\tilde{a}) \rangle \leq 0$$

for any $a, \tilde{a} \in \mathbb{R}^d$. Adding these inequalities, rearranging, and applying the Cauchy–Schwarz inequality give

$$\| P_A(a) - P_A(\tilde{a}) \|^2 \leq \langle a - \tilde{a}, P_A(a) - P_A(\tilde{a}) \rangle$$
$$\leq \| a - \tilde{a} \| \| P_A(a) - P_A(\tilde{a}) \| \tag{12}$$

Thus, $\| P_A(a) - P_A(\tilde{a}) \| \leq \| a - \tilde{a} \|$. That is, P_A is nonexpansive, and the inequality holds if and only if

$$P_A(a) - P_A(\tilde{a}) = c(a - \tilde{a}) \tag{13}$$

for some constant c. Likewise, P_B is nonexpansive. Therefore,

$$\| P_A[x - P_B(x - a)] - P_A[x - P_B(x - \tilde{a})] \|$$
$$\leq \| P_B(x - a) - P_B(x - \tilde{a}) \| \leq \| a - \tilde{a} \| \tag{14}$$

This proves that T is nonexpansive, hence continuous.

Next, we show that T is paracontractive. Suppose that $\tilde{a}$ is a fixed point, $a \neq \tilde{a}$, and equality holds throughout inequalities (14). Inequalities (12) and Equation (13) indicate that equality is achieved in the previous two inequalities only if

$$P_A[x - P_B(x - a)] - [x - P_B(x - a)]$$
$$= P_A[x - P_B(x - \tilde{a})] - [x - P_B(x - \tilde{a})]$$

and

$$P_B(x - a) - (x - a) = P_B(x - \tilde{a}) - (x - \tilde{a})$$

Subtracting the second of these equalities from the first gives

$$P_A[x - P_B(x - a)] - a = P_A[x - P_B(x - \tilde{a})] - \tilde{a} = 0$$

It follows that equality in inequalities (14) is achieved only if a is also a fixed point.

To show that T possesses a fixed point, note that given the closedness of $A + B$, there exists a closest point $\tilde{a} + \tilde{b}$ to x, where $\tilde{a} \in A$ and $\tilde{b} \in B$. Since block descent cannot improve the objective $f(a, b) = \frac{1}{2} \| x - a - b \|^2$ on the set $A \times B$ starting from $(\tilde{a}, \tilde{b})$, it is clear that $\tilde{a} = T(\tilde{a})$.

Finally, suppose that $\tilde{a}$ is any fixed point, and define $\tilde{b} = P_B(x - \tilde{a})$. To prove that $\tilde{a} + \tilde{b}$ minimizes the distance to x, it suffices to show that for every tangent vector $v = a + b - \tilde{a} - \tilde{b}$ at $\tilde{a} + \tilde{b}$, the directional derivative

$$d_v \frac{1}{2} \| x - \tilde{a} - \tilde{b} \|^2 = -\langle x - \tilde{a} - \tilde{b}, v \rangle$$
$$= -\langle x - \tilde{a} - \tilde{b}, a - \tilde{a} \rangle - \langle x - \tilde{a} - \tilde{b}, b - \tilde{b} \rangle$$

is nonnegative. However, the inequalities $-\langle x - \tilde{a} - \tilde{b}, a - \tilde{a} \rangle \geq 0$ and $-\langle x - \tilde{a} - \tilde{b}, b - \tilde{b} \rangle \geq 0$ hold because $\tilde{a}$ minimizes $a \mapsto \frac{1}{2} \| x - a - \tilde{b} \|^2$, and $\tilde{b}$ minimizes $b \mapsto \frac{1}{2} \| x - \tilde{a} - b \|^2$. Thus, any fixed point of T furnishes a minimum of the convex function $f(a, b)$ on the set $A \times B$.

4 Bregman Majorization

Bregman majorization is a technique for constructing a sequence of surrogate functions pertinent to an MM algorithm. Let us first define the notion of Bregman divergence.

Definition 11. (Bregman divergence). *For a proper convex function $\phi(x)$ that is continuously differentiable on* **int** dom ϕ, *the Bregman divergence $B_\phi : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is defined as*

$$B_\phi(x \| y) = \phi(x) - \phi(y) - \langle \nabla\phi(y), x - y \rangle, \quad x, y \in \textbf{int} \text{ dom } \phi$$

We are concerned with the following optimization problem:

$$\min_{x \in C} f(x), \quad C \subset \mathcal{X} \text{ is closed and convex} \tag{15}$$

where $f(x)$ is convex, proper, and lower semicontinuous. In order to solve this problem, the Bregman majorization method constructs the sequence of surrogate functions

$$g(x \mid x_n) = f(x) + B_\phi(x \| x_n)$$

and successively minimizes these. This is a valid MM algorithm since the following properties of the Bregman divergence are immediate from definition:

1. $B_\phi(x \| y) \geq 0$;
2. $B_\phi(x \| x) = 0$;
3. If ϕ is strictly convex, then $B_\phi(x \| y) = 0$ if and only if $x = y$.

Thus, $g(x \mid x_n) \geq f(x)$ for all x and $g(x_n \mid x_n) = f(x_n)$. We can choose $\phi(x)$ so that for **cl** dom $\phi = C$.

The subsequent section studies the convergence property of the Bregman majorization.

4.1 Convergence Analysis via SUMMA

The sequential unconstrained minimization method algorithm (SUMMA)[53] is a class of algorithms for solving optimization problems of the form

$$\min_{x \in C} f(x), \quad C \subset \mathcal{X} \text{ is closed} \tag{16}$$

by minimizing a sequence of auxiliary functions

$$G_n(x) = f(x) + g_{n+1}(x), \quad n = 1, 2, \ldots$$

over $\mathcal{X}$. The minimizer of $G_n(x)$ is denoted by x_n. The conditions imposed on the sequence of functions $g_n(x)$ are:

1. $g_n(x) \geq 0$ for all $x \in \mathcal{X}$;
2. $g_n(x_{n-1}) = 0$;
3. $G_n(x) - G_n(x_n) \geq g_{n+1}(x)$ for all $x \in C$.

If $g_n(x)$ depends on n only through the iterate x_n, then this method coincides with the MM algorithm by identifying $G_n(x) = g(x \mid x_n)$ and $g_n(x) = g(x \mid x_{n-1}) - f(x)$, with the additional requirement

$$g(\boldsymbol{x} \mid \boldsymbol{x}_n) - g(\boldsymbol{x}_{n+1} \mid \boldsymbol{x}_n) \geq g(\boldsymbol{x} \mid \boldsymbol{x}_{n+1}) - f(\boldsymbol{x}) \tag{17}$$

for all $\boldsymbol{x} \in C$.

Let us show that condition (17) is satisfied by the Bregman majorization $g(\boldsymbol{x} \mid \boldsymbol{x}_n) = f(\boldsymbol{x}) + \phi(\boldsymbol{x}) - \phi(\boldsymbol{x}_n) - \langle \nabla\phi(\boldsymbol{x}_n), \boldsymbol{x} - \boldsymbol{x}_n \rangle$. The optimality condition for minimizing $g(\boldsymbol{x} \mid \boldsymbol{x}_n)$ is

$$\boldsymbol{0} \in \partial f(\boldsymbol{x}_{n+1}) + \nabla\phi(\boldsymbol{x}_{n+1}) - \nabla\phi(\boldsymbol{x}_n)$$

For the appropriate choice of $\boldsymbol{s}_{n+1} \in \partial f(\boldsymbol{x}_{n+1})$, it follows that

$$\begin{aligned}
g(\boldsymbol{x} \mid \boldsymbol{x}_n) - g(\boldsymbol{x}_{n+1} \mid \boldsymbol{x}_n) &= f(\boldsymbol{x}) - f(\boldsymbol{x}_{n+1}) + \phi(\boldsymbol{x}) - \phi(\boldsymbol{x}_{n+1}) \\
&\quad - \langle \nabla\phi(\boldsymbol{x}_n), \boldsymbol{x} - \boldsymbol{x}_{n+1} \rangle \\
&= f(\boldsymbol{x}) - f(\boldsymbol{x}_{n+1}) - \langle \boldsymbol{s}_{n+1}, \boldsymbol{x} - \boldsymbol{x}_{n+1} \rangle \\
&\quad + \phi(\boldsymbol{x}) - \phi(\boldsymbol{x}_{n+1}) - \langle \nabla\phi(\boldsymbol{x}_{n+1}), \boldsymbol{x} - \boldsymbol{x}_{n+1} \rangle \\
&\geq B_\phi(\boldsymbol{x} \parallel \boldsymbol{x}_{n+1}) = g(\boldsymbol{x} \mid \boldsymbol{x}_{n+1}) - f(\boldsymbol{x})
\end{aligned}$$

where the last inequality is a consequence of the convexity of $f(\boldsymbol{x})$.

The following propositions concern convergence of MM algorithms satisfying condition (17).

Proposition 14. *Assume (a) $p^* = \inf_{x \in C} f(\boldsymbol{x}) > -\infty$ and (b) $\boldsymbol{x}_n \in C$ for each n. If condition (17) holds, then any MM sequence generated by the map $\boldsymbol{x}_{n+1} \in \operatorname{argmin}_{x \in \mathcal{X}} g(\boldsymbol{x} \mid \boldsymbol{x}_n)$ satisfies $\lim_{n \to \infty} f(\boldsymbol{x}_n) = p^*$.*

Proof. By the descent property of MM and the bound $f(\boldsymbol{x}_n) \geq p^* > -\infty$ given $\boldsymbol{x}_n \in C$, the sequence $f(\boldsymbol{x}_n)$ converges to a limit $d \geq p^*$. Suppose for some $\boldsymbol{x} \in C$ that $f(\boldsymbol{x}) < d$. Then, by condition (17),

$$\begin{aligned}
[g(\boldsymbol{x} \mid \boldsymbol{x}_n) - f(\boldsymbol{x})] - [g(\boldsymbol{x} \mid \boldsymbol{x}_{n+1}) - f(\boldsymbol{x})] &\geq g(\boldsymbol{x}_{n+1} \mid \boldsymbol{x}_n) - f(\boldsymbol{x}) \\
&\geq f(\boldsymbol{x}_{n+1}) - f(\boldsymbol{x}) \\
&\geq d - f(\boldsymbol{x}) \\
&> 0
\end{aligned}$$

Thus, the sequence $g(\boldsymbol{x} \mid \boldsymbol{x}_n) - f(\boldsymbol{x})$ decreases, and its successive differences are bounded away from zero. The latter property contradicts the requirement for the surrogate function that $g(\boldsymbol{x} \mid \boldsymbol{x}_n) \geq f(\boldsymbol{x})$, and therefore $d = p^*$.

Proposition 15. *In addition to the assumptions of Proposition 14, further assume that (c) the minimum p^* is attained and the set F of the minimizers of $f(\boldsymbol{x})$ in C is nonempty, (d) $f(\boldsymbol{x})$ is continuous on $D \subset \mathcal{X}$ such that $\mathbf{cl}D = C$, (e) for each n, $g(\boldsymbol{x} \mid \boldsymbol{x}_n)$ is μ-strongly convex with respect to the norm $\|\cdot\|$ and $\operatorname{dom}g(\cdot \mid \boldsymbol{x}_n) = D$, and (f) $g(\boldsymbol{x} \mid \boldsymbol{x}_n) - f(\boldsymbol{x}) \leq \frac{L}{2} \| \boldsymbol{x} - \boldsymbol{x}_n \|^2$ for all $\boldsymbol{x} \in D$ and each n. If condition (17) holds, then the MM sequence $\boldsymbol{x}_{n+1} = \operatorname{argmin}_{x \in \mathcal{X}} g(\boldsymbol{x} \mid \boldsymbol{x}_n)$ converges to a point in F.*

Proof. Because of strong convexity, the minimum of $g(\boldsymbol{x} \mid \boldsymbol{x}_n)$ is uniquely attained for each n. Furthermore, for any $\boldsymbol{x} \in D$,

$$g(\boldsymbol{x} \mid \boldsymbol{x}_n) - g(\boldsymbol{x}_{n+1} \mid \boldsymbol{x}_n) \geq \frac{\mu}{2} \| \boldsymbol{x} - \boldsymbol{x}_{n+1} \|^2 \tag{18}$$

Let $\boldsymbol{y} \in F$ be a minimizer of $f(\boldsymbol{x})$ in C. Since $f(\boldsymbol{x}_{n+1}) \leq g(\boldsymbol{x}_{n+1} \mid \boldsymbol{x}_n)$,

$$g(\boldsymbol{y} \mid \boldsymbol{x}_n) - f(\boldsymbol{x}_{n+1}) \geq g(\boldsymbol{y} \mid \boldsymbol{x}_n) - g(\boldsymbol{x}_{n+1} \mid \boldsymbol{x}_n) \geq \frac{\mu}{2} \| \boldsymbol{y} - \boldsymbol{x}_{n+1} \|^2 \tag{19}$$

where the last inequality follows from the strong convexity of $g(x \mid x_n)$. Condition (17) also implies

$$[g(y \mid x_n) - f(y)] - [g(y \mid x_{n+1}) - f(y)] \geq g(x_{n+1} \mid x_n) - f(y)$$
$$\geq f(x_{n+1}) - p^* \geq 0$$

Hence, the decreasing nonnegative sequence $g(y \mid x_n) - f(y)$ has a limit. In addition, $f(y) - f(x_{n+1})$ tends to zero by Proposition 14. It follows that the leftmost side of inequality (19) tends to a limit, and the sequence x_n is bounded.

Suppose that the convergent subsequence x_{n_m} of x_n has a limit z. By continuity, $f(z) = \lim_{m \to \infty} f(x_{n_m}) = p^*$, so z is also optimal. Now,

$$0 \leq g(z \mid x_n) - g(x_{n+1} \mid x_n)$$
$$= [g(z \mid x_n) - f(z)] + f(z) - f(x_{n+1}) - [g(x_{n+1} \mid x_n) - f(x_{n+1})]$$
$$\leq g(z \mid x_n) - f(z)$$
$$\leq \frac{L}{2} \parallel x_n - z \parallel^2$$

due to $f(z) \leq f(x_{n+1})$, $g(x_{n+1} \mid x_n) - f(x_{n+1}) \geq 0$, and assumption (f). Again by Condition (17), we further have

$$0 \leq g(z \mid x_n) - g(x_{n+1} \mid x_n) \leq g(z \mid x_n) - f(z) \leq g(z \mid x_{n-1}) - g(x_n \mid x_{n-1}) \tag{20}$$

Thus, the nonnegative sequence $g(z \mid x_n) - g(x_{n+1} \mid x_n)$ is monotonically decreasing and convergent. Its subsequence $g(z \mid x_{n_m}) - g(x_{n_m+1} \mid x_{n_m})$ is also bounded by $\frac{L}{2} \parallel x_{n_m} - z \parallel^2$, which converges to zero. Thus, the whole sequence tends to zero. By inequality (20), it follows that the sequence $g(z \mid x_n) - f(z)$ converges to zero.

The final inequality

$$g(z \mid x_n) - f(z) = g(z \mid x_n) - g(x_{n+1} \mid x_n) + g(x_{n+1} \mid x_n) - f(z)$$
$$\geq \frac{\mu}{2} \parallel z - x_{n+1} \parallel^2 + f(x_{n+1}) - f(z)$$

now proves that the entire sequence x_n converges to $z \in F$.

Remark 1. Assumption (e) (uniform strong convexity of the surrogate functions) is much less restrictive than assuming strong convexity on the objective $f(x)$. For example, assumption (e) is satisfied when $f(x)$ is convex, and the convex function $\phi(x)$ defining the Bregman divergence is μ-strongly convex.

Remark 2. Assumption (f) is satisfied if $\phi(x)$ is L-smooth. Assumption (f) can be replaced by

(f′) $g(x \mid y)$ is continuous in y in D.

This is the condition implicitly imposed in the proof of Proposition 7.4.1 in Lange[41]. (This assumption is not made perfectly clear in the statement of the proposition.) Assumption (f′) is satisfied, when $\phi(x)$ is a Bregman−Legendre function[53, 54].

4.2 Examples

4.2.1 Proximal gradient method

The proximal gradient method minimizes $f(\boldsymbol{x}) = f_0(\boldsymbol{x}) + h(\boldsymbol{x})$ over $C = \mathcal{X}$, where both $f_0(\boldsymbol{x})$ and $h(\boldsymbol{x})$ are convex, proper, and lower semicontinuous. It is further assumed that $f_0(\boldsymbol{x})$ is L-smooth. The algorithm iteratively solves

$$\boldsymbol{x}_{n+1} = \operatorname{argmin}_{\boldsymbol{x}} \left\{ f_0(\boldsymbol{x}_n) + \langle \nabla f_0(\boldsymbol{x}_n), \boldsymbol{x} - \boldsymbol{x}_n \rangle + h(\boldsymbol{x}) + \frac{1}{2\alpha} \| \boldsymbol{x} - \boldsymbol{x}_n \|^2 \right\} \tag{21}$$

for a step size $0 < \alpha < L^{-1}$. To see that the proximal gradient algorithm is an instance of Bregman majorization, set $\phi(\boldsymbol{x}) = \frac{1}{2\alpha} \| \boldsymbol{x} \|^2 - f_0(\boldsymbol{x})$. Then,

$$\begin{aligned} f(\boldsymbol{x}) + B_\phi(\boldsymbol{x} \parallel \boldsymbol{x}_n) &= f_0(\boldsymbol{x}) + h(\boldsymbol{x}) + \frac{1}{2\alpha} \| \boldsymbol{x} \|^2 - f_0(\boldsymbol{x}) - \frac{1}{2\alpha} \| \boldsymbol{x}_n \|^2 + f_0(\boldsymbol{x}_n) \\ &\quad - \langle \frac{1}{\alpha} \boldsymbol{x}_n - \nabla f_0(\boldsymbol{x}_n), \boldsymbol{x} - \boldsymbol{x}_n \rangle \\ &= f_0(\boldsymbol{x}_n) + \langle \nabla f_0(\boldsymbol{x}_n), \boldsymbol{x} - \boldsymbol{x}_n \rangle + h(\boldsymbol{x}) + \frac{1}{2\alpha} \| \boldsymbol{x} - \boldsymbol{x}_n \|^2 \end{aligned} \tag{22}$$

as desired. It remains to verify that $f(\boldsymbol{x})$ and $\phi(\boldsymbol{x})$ satisfy conditions (a) through (f) of Propositions 14 and 15. Conditions (a) and (c) are assumed; (b) and (d) are true. Condition (e) is satisfied since $\alpha \in (0, 1/L)$. The following fact is well known:

Lemma 2. *A differentiable convex function $f(\boldsymbol{x})$ is L-smooth $\nabla f(\boldsymbol{x})$ if and only if $\frac{L}{2} \|\boldsymbol{x}\|^2 - f(\boldsymbol{x})$ is convex.*

Then, since $\phi(\boldsymbol{x}) = \frac{1}{2} \left(\frac{1}{\alpha} - L \right) \|\boldsymbol{x}\|^2 + \frac{L}{2} \|\boldsymbol{x}\|^2 - f(\boldsymbol{x})$ and $\frac{1}{\alpha} > L$, ϕ is $(\frac{1}{\alpha} - L)$-strongly convex. To check condition (f), we invoke the Baillon–Haddad theorem:

Lemma 3. *If function $f(\boldsymbol{x})$ is convex, differentiable, and L-smooth, then*

$$\langle \nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle \geq \frac{1}{L} \| \nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}) \|^2$$

Note $\nabla \phi(\boldsymbol{x}) = \frac{1}{\alpha} \boldsymbol{x} - \nabla f_0(\boldsymbol{x})$. Then,

$$\begin{aligned} \| \nabla \phi(\boldsymbol{x}) - \nabla \phi(\boldsymbol{y}) \|^2 &= \| \alpha^{-1}(\boldsymbol{x} - \boldsymbol{y}) - [\nabla f_0(\boldsymbol{x}) - \nabla f_0(\boldsymbol{y})] \|^2 \\ &= \frac{1}{\alpha^2} \| \boldsymbol{x} - \boldsymbol{y} \|^2 + \| \nabla f_0(\boldsymbol{x}) - \nabla f_0(\boldsymbol{y}) \|^2 - \frac{2}{\alpha} \langle \boldsymbol{x} - \boldsymbol{y}, \nabla f_0(\boldsymbol{x}) - \nabla f_0(\boldsymbol{y}) \rangle \\ &\leq \frac{1}{\alpha^2} \| \boldsymbol{x} - \boldsymbol{y} \|^2 + \| \nabla f_0(\boldsymbol{x}) - \nabla f_0(\boldsymbol{y}) \|^2 - \frac{2}{\alpha L} \| \nabla f_0(\boldsymbol{x}) - \nabla f_0(\boldsymbol{y}) \|^2 \\ &\leq \frac{1}{\alpha^2} \| \boldsymbol{x} - \boldsymbol{y} \|^2. \end{aligned}$$

The first inequality is due to Lemma 3. The last inequality holds since $\alpha \in (0, 1/L)$ implies $1 - \frac{2}{\alpha L} \leq 0$. Therefore, $\nabla \phi(\boldsymbol{x})$ is $1/\alpha$-Lipschitz continuous and condition (f) is satisfied.

We summarize the discussion above as follows:

Proposition 16. *Suppose that $f_0(\boldsymbol{x})$ and $h(\boldsymbol{x})$ are convex, proper, and lower semicontinuous. If $f_0(\boldsymbol{x})$ is L-smooth, then for $0 < \alpha < 1/L$, the proximal gradient iteration (21) converges to a minimizer of $f(\boldsymbol{x}) = f_0(\boldsymbol{x}) + h(\boldsymbol{x})$ if it exists.*

Remark 3. Lemma 3 suggests that $\nabla\phi$ is $1/\alpha$-Lipschitz continuous if $0 < \alpha < 2/L$; in other words, the step size may be doubled. Indeed, employing monotone operator theory[55, 56] it can be shown that iteration (21) converges for $1/L \leq \alpha < 2/L$ as well. Even though the MM interpretation is lost for this range of step size, the descent property remains intact[57, 58].

Remark 4. The assumption that $h(\boldsymbol{x})$ is convex can be relaxed: if $h(\boldsymbol{x})$ is ρ-weakly convex, which means $h(\boldsymbol{x}) + \frac{\rho}{2} \| \boldsymbol{x}\|^2$ is convex, and $f_0(\boldsymbol{x})$ is ρ-strongly convex as well as L-smooth (this implies $\rho \leq L$), then the objective $f(\boldsymbol{x})$ remains convex. The inner optimization problem in iteration (21) is also strongly convex if $\rho\alpha < 1$ and $\boldsymbol{x}_{n+1}$ is uniquely determined. The latter condition is guaranteed if $\alpha \in (0, 1/L)$, and the conclusion of Proposition 16 holds. In fact, using monotone operator theory, a larger step size $\alpha \in (0, \frac{2}{L+\rho})$ is allowed[58]. Statistical applications include nonconvex sparsity-inducing penalties such as the MCP[59].

4.2.2 Mirror descent method

For the constrained problem (16) and the Euclidean norm $\|\cdot\|_2$, the proximal gradient method takes the form of *projected gradient*

$$
\begin{aligned}
\boldsymbol{x}_{n+1} &= \operatorname{argmin}_{\boldsymbol{x}\in C} \left\{ f(\boldsymbol{x}_n) + \langle \nabla f(\boldsymbol{x}_n), \boldsymbol{x} - \boldsymbol{x}_n \rangle + \frac{1}{2\alpha} \| \boldsymbol{x} - \boldsymbol{x}_n\|_2^2 \right\} \\
&= P_C(\boldsymbol{x}_n - \alpha\nabla f(\boldsymbol{x}_n))
\end{aligned}
\tag{23}
$$

This method relies heavily on the Euclidean geometry of $\mathbb{R}^d$, not C: $\|\cdot\|_2 = \langle\cdot,\cdot\rangle$. If the distance measure $\frac{1}{2} \| \mathbf{x} - \mathbf{y} \|_2^2$ is replaced by something else (say $d(\mathbf{x}, \mathbf{y})$) that better reflects the geometry of C, then update such as

$$
\boldsymbol{x}_{n+1} = P_C^d \left(\arg\min_{\boldsymbol{x}\in\mathbb{R}^d} \left\{ f(\boldsymbol{x}_n) + \langle \nabla f(\boldsymbol{x}_n), \boldsymbol{x} - \boldsymbol{x}_n \rangle + \frac{1}{\alpha}d(\boldsymbol{x}, \boldsymbol{x}_n) \right\} \right)
\tag{24}
$$

may converge faster. Here,

$$
P_C^d(\mathbf{y}) = \operatorname{argmin}_{\boldsymbol{x}\in C}d(\mathbf{x}, \mathbf{y})
$$

is a new (non-Euclidean) projection operator that reflects the geometry of C.

To see that iteration (24) is a Bregman majorization for an appropriately chosen $d(\cdot,\cdot)$, let

$$
d(\boldsymbol{x}, \boldsymbol{y}) = B_\psi(\boldsymbol{x} \| \boldsymbol{y}) = \psi(\boldsymbol{x}) - \psi(\boldsymbol{y}) - \langle \nabla\psi(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle \geq \frac{1}{2} \| \boldsymbol{x} - \boldsymbol{y}\|^2
$$

for a 1-strongly convex (with respect to some norm $\| \cdot \|$) and continuously differentiable function ψ in C, and set $\phi(\boldsymbol{x}) = \frac{1}{\alpha}\psi(\boldsymbol{x}) - f(\boldsymbol{x})$. Similar to Equation (22), we have

$$
\begin{aligned}
f(\boldsymbol{x}) + B_\phi(\boldsymbol{x} \| \boldsymbol{x}_n) &= f(\boldsymbol{x}_n) + \langle \nabla f(\boldsymbol{x}_n), \boldsymbol{x} - \boldsymbol{x}_n \rangle \\
&\quad + \frac{1}{\alpha}[\psi(\boldsymbol{x}) - \psi(\boldsymbol{x}_n) - \langle \nabla\psi(\boldsymbol{x}_n), \boldsymbol{x} - \boldsymbol{x}_n \rangle] \\
&= f(\boldsymbol{x}_n) + \langle \nabla f(\boldsymbol{x}_n), \boldsymbol{x} - \boldsymbol{x}_n \rangle + \frac{1}{\alpha}d(\boldsymbol{x}, \boldsymbol{x}_n)
\end{aligned}
$$

Let $\tilde{\boldsymbol{x}}_{n+1}$ be the *unconstrained* minimizer of $f(\boldsymbol{x}) + B_\phi(\boldsymbol{x} \| \boldsymbol{x}_n)$ (which is unique since $d(\boldsymbol{x}, \boldsymbol{x}_n)$ is strongly convex in $\boldsymbol{x}$). The associated optimality condition entails

$$
\nabla\psi(\tilde{\boldsymbol{x}}_{n+1}) = \nabla\psi(\boldsymbol{x}_n) - \alpha\nabla f(\boldsymbol{x}_n)
\tag{25}
$$

Then,

$$
\begin{aligned}
\boldsymbol{x}_{n+1} &= \operatorname{argmin}_{\boldsymbol{x} \in C} d(\boldsymbol{x}, \tilde{\boldsymbol{x}}_{n+1}) \\
&= \operatorname{argmin}_{\boldsymbol{x} \in C} \{ \psi(\boldsymbol{x}) - \psi(\tilde{\boldsymbol{x}}_{n+1}) - \langle \nabla \psi(\tilde{\boldsymbol{x}}_{n+1}), \boldsymbol{x} - \tilde{\boldsymbol{x}}_{n+1} \rangle \} \\
&= \operatorname{argmin}_{\boldsymbol{x} \in C} \{ \psi(\boldsymbol{x}) - \langle \nabla \psi(\tilde{\boldsymbol{x}}_{n+1}), \boldsymbol{x} \rangle \} \\
&= \operatorname{argmin}_{\boldsymbol{x} \in C} \{ \psi(\boldsymbol{x}) - \langle \nabla \psi(\boldsymbol{x}_n) - \alpha \nabla f(\boldsymbol{x}_n), \boldsymbol{x} - \boldsymbol{x}_n \rangle \} \\
&= \operatorname{argmin}_{\boldsymbol{x} \in C} \{ f(\boldsymbol{x}) + \phi(\boldsymbol{x}) - \langle \nabla \phi(\boldsymbol{x}_n), \boldsymbol{x} - \boldsymbol{x}_n \rangle - \phi(\boldsymbol{x}_n) \} \\
&= \operatorname{argmin}_{\boldsymbol{x} \in C} \{ f(\boldsymbol{x}) + B_\phi(\boldsymbol{x} \parallel \boldsymbol{x}_n) \}
\end{aligned}
$$

as sought. To establish iterate convergence via SUMMA, we see that just as the proximal gradient method, $f(\boldsymbol{x})$ and $\phi(\boldsymbol{x})$ satisfy conditions (a) through (e) of Propositions 14 and 15 if f is L-smooth and $\alpha \in (0, 1/L)$. In particular,

$$
\phi(\boldsymbol{x}) = \frac{1}{\alpha} \psi(\boldsymbol{x}) - f(\boldsymbol{x}) \geq \frac{1}{2\alpha} \parallel \boldsymbol{x} \parallel^2 - f(\boldsymbol{x})
$$

to check condition (e). Condition (f') is fulfilled since $B_\phi(\boldsymbol{x} \parallel \boldsymbol{y}) = \phi(\boldsymbol{x}) - \phi(\boldsymbol{y}) - \langle \nabla \phi(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle$ is continuous in $\boldsymbol{y}$ by construction.

Computation of $\boldsymbol{x}_{n+1}$ can be further analyzed. It is well known that if ψ is μ-strongly convex, then ψ^* is $1/\mu$-smooth, where ψ^* is the Fenchel conjugate function of ψ[55]:

$$
\psi^*(\boldsymbol{y}) = \sup_{\boldsymbol{x} \in \operatorname{dom} \psi} \langle \boldsymbol{x}, \boldsymbol{y} \rangle - \phi(\boldsymbol{x})
$$

Hence, $\nabla \psi^*$ is well defined. Furthermore, $\nabla \psi^*(\nabla \psi(\boldsymbol{x})) = \boldsymbol{x}$. Therefore, the unconstrained optimality condition (25) is equivalent to

$$
\tilde{\boldsymbol{x}}_{n+1} = \nabla \psi^*(\nabla \psi(\boldsymbol{x}_n) - \alpha \nabla f(\boldsymbol{x}_n))
$$

and we decompose the update (24) into three steps:

$$
\begin{aligned}
\boldsymbol{y}_{n+1} &= \nabla \psi(\boldsymbol{x}_n) - \alpha \nabla f(\boldsymbol{x}_n) \quad \text{(gradient step)} \\
\tilde{\boldsymbol{x}}_{n+1} &= \nabla \psi^*(\boldsymbol{y}_{n+1}) \quad \text{(mirroring step)} \\
\boldsymbol{x}_{n+1} &= P_C^d(\tilde{\boldsymbol{x}}_{n+1}) \quad \text{(projection step)}
\end{aligned}
$$

Hence, Bregman majorization with $\phi(\boldsymbol{x}) = \frac{1}{\alpha} \psi(\boldsymbol{x}) - f(\boldsymbol{x})$ coincides with the mirror descent method under B_ψ[60]. The first step performs the gradient descent step in the dual space $\mathcal{X}^*$ of $\mathcal{X}$, and the second step maps the dual vector back to the primal space by the inverse mapping $\nabla \psi^* = (\nabla \psi)^{-1}$. The final step projects (in a non-Euclidean fashion) the mapped primal vector onto the constraint set C.

Example 3. (Exponentiated gradient). As a concrete instance of mirror descent, consider optimization over probability simplex $C = \Delta^{d-1} = \{ \mathbf{x} \in \mathcal{X} = \mathbb{R}^d : \sum_{i=1}^d x_i = 1, \ x_i \geq 0, \ i = 1, \dots, d \}$. An appropriate Bregman divergence is the Kullback–Leibler divergence, that is, we use negative entropy $\psi(\boldsymbol{x}) = \sum_{i=1}^d x_i \log x_i - \sum_{i=1}^d x_i$. It is easy to check, using the Taylor expansion and the Cauchy–Schwarz inequality, that ψ is 1-strongly convex with respect to the ℓ_1 norm $\parallel \boldsymbol{x} \parallel_1 = \sum_{i=1}^d |x_i|$ within C. Furthermore, we have $\nabla \psi(\boldsymbol{x}) = (\log x_1, \dots, \log x_d)^T =: \log \boldsymbol{x}$ and $\nabla \psi^*(\boldsymbol{y}) = (\nabla \phi)^{-1}(\boldsymbol{y}) = (e^{y_1}, \dots, e^{y_d})^T =: \exp(\boldsymbol{y})$. The mirror descent or Bregman MM update is then

$$
\begin{aligned}
\boldsymbol{y}_{n+1} &= \log \boldsymbol{x}_n - \alpha \nabla f(\boldsymbol{x}_n) \\
\tilde{\boldsymbol{x}}_{n+1} &= \exp(\boldsymbol{y}_n) = \boldsymbol{x}_n \odot \exp(-\alpha \nabla f(\boldsymbol{x}_n)) \\
\boldsymbol{x}_{n+1} &= \tilde{\boldsymbol{x}}_{n+1} / Z_t
\end{aligned}
$$

where $\odot$ denotes an elementwise product, and

$$Z_t = \sum_{i=1}^{d} x_{n,i} \exp(-\alpha \nabla f(\boldsymbol{x}_n)_i)$$

is the normalization constant. The last step is because

$$P_C^d(\boldsymbol{y}) = \operatorname{argmin}_{\boldsymbol{x} \in \Delta^{d-1}} B_\psi(\boldsymbol{x} \parallel \boldsymbol{y})$$

$$= \operatorname{argmin}_{x_i \geq 0, \sum_{i=1}^{d} x_i = 1} \sum_{i=1}^{d} \left(x_i \log \frac{x_i}{y_i} - x_i + y_i \right)$$

$$= \operatorname{argmin}_{x_i \geq 0, \sum_{i=1}^{d} x_i = 1} \sum_{i=1}^{d} \left(x_i \log \frac{x_i}{y_i} \right)$$

and the associated Lagrangian

$$\mathcal{L}(\boldsymbol{x}, \mu) = \sum_{i=1}^{d} \left(x_i \log \frac{x_i}{y_i} \right) + \mu \left(\sum_{i=1}^{d} x_i - 1 \right)$$

yields

$$x_i = y_i \exp(-\mu - 1) = c y_i, \quad i = 1, \dots, d$$

for some $c > 0$. Summing these over all i yields $c = 1/(\sum_{i=1}^{d} y_i)$ to have

$$x_i = \frac{y_i}{\sum_{j=1}^{d} y_j}, \quad i = 1, \dots, d$$

This special case is called the *exponentiated gradient method*[61, 62].

Related Articles

EM Algorithm; **Non-Convex Optimization**; **Optimization and Nonlinear Equations**; **Constrained Optimization**; **Optimization Methods**; **Optimization in Statistics**

References

[1] Becker, M.P., Yang, I., and Lange, K. (1997) EM algorithms without missing data. *Stat. Methods Med. Res.*, **6** (1), 38–54.
[2] Lange, K., Hunter, D.R., and Yang, I. (2000) Optimization transfer using surrogate objective functions. *J. Comput. Graph. Statist.*, **9** (1), 1–59. [With discussion, and a rejoinder by Hunter and Lange].
[3] Hunter, D.R. and Lange, K. (2004) A tutorial on MM algorithms. *Am. Stat.*, **58**, 30–37.
[4] Borg, I. and Groenen, P.J.F. (2005) *Modern Multidimensional Scaling: Theory and Applications*, Springer Series in Statistics, 2nd edn, Springer, New York.
[5] Hunter, D.R. and Lange, K. (2000) Quantile regression via an MM algorithm. *J. Comput. Graph. Statist.*, **9** (1), 60–77.
[6] Hunter, D.R. (2004) MM algorithms for generalized Bradley-Terry models. *Ann. Statist.*, **32** (1), 384–406.
[7] Hunter, D.R. and Li, R. (2005) Variable selection using MM algorithms. *Ann. Statist.*, **33** (4), 1617–1642.
[8] Yen, T.-J. (2011) A majorization-minimization approach to variable selection using spike and slab priors. *Ann. Statist.*, **39** (3), 1748–1775.
[9] Bien, J. and Tibshirani, R.J. (2011) Sparse estimation of a covariance matrix. *Biometrika*, **98** (4), 807–820.

[10] Lee, S. and Huang, J.Z. (2013) A coordinate descent MM algorithm for fast computation of sparse logistic PCA. *Comput. Statist. Data Anal.*, **62**, 26–38.

[11] Zhou, H. and Lange, K. (2010) MM algorithms for some discrete multivariate distributions. *J. Comput. Graph. Stat.*, **19**, 645–665.

[12] Zhang, Y., Zhou, H., Zhou, J., and Sun, W. (2017) Regression models for multivariate count data. *J. Comput. Graph. Stat.*, **26** (1), 1–13.

[13] Zhou, H., Hu, L., Zhou, J., and Lange, K. (2019) MM algorithms for variance components models. *J. Comput. Graph. Statist.*, **28** (2), 350–361.

[14] Sun, Y., Babu, P., and Palomar, D.P. (2015) Regularized robust estimation of mean and covariance matrix under heavy-tailed distributions. *IEEE Trans. Signal Process.*, **63** (12), 3096–3109.

[15] Hunter, D.R. and Lange, K. (2002) Computing estimates in the proportional odds model. *Ann. Inst. Statist. Math.*, **54** (1), 155–168.

[16] Ding, J., Tian, G.-L. and Yuen, K.C. (2015) A new MM algorithm for constrained estimation in the proportional hazards model. *Comput. Statist. Data Anal.*, **84**, 135–151.

[17] Lange, K. and Zhou, H. (2014) MM algorithms for geometric and signomial programming. *Math. Program. Series A*, **143**, 339–356.

[18] Chi, E.C., Zhou, H. and Lange, K. (2014) Distance majorization and its applications. *Math. Program.*, **146** (1-2), 409–436.

[19] Xu, J., Chi, E. and Lange, K. (2017) Generalized linear model regression under distance-to-set penalties, in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., Red Hook, NY, pp. 1385–1395.

[20] Keys, K.L., Zhou, H., and Lange, K. (2019) Proximal distance algorithms: theory and examples. *J. Mach. Learn. Res.*, **20** (66), 1–38.

[21] Lange, K. and Carson, R. (1984) EM reconstruction algorithms for emission and transmission tomography. *J. Comput. Assist. Tomogr.*, **8** (2), 306–316.

[22] Figueiredo, M.A.T., Bioucas-Dias, J.M., and Nowak, R.D. (2007) Majorization–minimization algorithms for wavelet-based image restoration. *IEEE Trans. Image Process.*, **16** (12), 2980–2991.

[23] Lee, D.D. and Seung, H.S. (1999) Learning the parts of objects by non-negative matrix factorization. *Nature*, **401** (6755), 788–791.

[24] Mazumder, R., Hastie, T., and Tibshirani, R. (2010) Spectral regularization algorithms for learning large incomplete matrices. *J. Mach. Learn. Res.*, **11**, 2287–2322.

[25] Chi, E.C., Zhou, H., Chen, G.K. *et al.* (2013) Genotype imputation via matrix completion. *Genome Res.*, **23** (3), 509–518.

[26] Chi, E. and Lange, K. (2015) Splitting methods for convex clustering. *J. Comput. Graph. Stat.*, **24** (4), 994–1013.

[27] Xu, J. and Lange, K. (2019) By all means, *k*-means, *under review*.

[28] Wu, T.T. and Lange, K. (2010) Multicategory vertex discriminant analysis for high-dimensional data. *Ann. Appl. Stat.*, **4** (4), 1698–1721.

[29] Nguyen, H.D. (2017) An introduction to majorization-minimization algorithms for machine learning and statistical estimation. *WIREs Data Min. Knowl. Discov.*, **7** (2), e1198.

[30] Lange, K. (2016) *MM Optimization Algorithms*, Society for Industrial and Applied Mathematics, Philadelphia, PA.

[31] Sun, Y., Babu, P., and Palomar, D.P. (2017) Majorization-minimization algorithms in signal processing, communications, and machine learning. *IEEE Trans. Signal Process.*, **65** (3), 794–816.

[32] Nguyen, H.D. (2017) An introduction to Majorization-Minimization algorithms for machine learning and statistical estimation. *WIREs Data Min. Knowl. Discov.*, **7** (2), e1198.

[33] Zhou, H. and Zhang, Y. (2012) EM vs MM: a case study. *Comput. Stat. Data Anal.*, 3909–3920. **56**:

[34] Zangwill, W.I. and Mond, B. (1969) *Nonlinear Programming: A Unified Approach*, Prentice-Hall International Series in Management, Prentice-Hall Inc., Englewood Cliffs, N.J.

[35] Luenberger, D.G. and Ye, Y. (2008) Linear and Nonlinear Programming, in *International Series in Operations Research & Management Science*, vol. 116, 3rd edn, Springer, New York.

[36] Vaida, F. (2005) Parameter convergence for EM and MM algorithms. *Stat. Sin.*, **15**, 831–840.

[37] van de Geer, J.P. (1984) Linear relations among *k* sets of variables. *Psychometrika*, **49** (1), 79–94.

[38] Ten Berge, J.M.F. (1988) Generalized approaches to the maxbet problem and the maxdiff problem, with applications to canonical correlations. *Psychometrika*, **53** (4), 487–494.

[39] Hanafi, M. and Kiers, H.A. (2006) Analysis of k sets of data, with differential emphasis on agreement between and within sets. *Comput. Stat. Data Anal.*, **51** (3), 1491–1508.

[40] Ten Berge, J.M.F. and Knol, D.L. (1984) Orthogonal rotations to maximal agreement for two or more matrices of different column orders. *Psychometrika*, **49** (1), 49–55.

[41] Lange, K. (2016) MM Optimization Algorithms, SIAM.

[42] Won, J.-H., Zhou, H., and Lange, K. (2018) Orthogonal trace-sum maximization: applications, local algorithms, and global optimality, *arXiv preprint arXiv:1811.03521*.

[43] Ten Berge, J.M.F. (1977) Orthogonal procrustes rotation for two or more matrices. *Psychometrika*, **42** (2), 267–276.

[44] Absil, P.-A. and Malick, J. (2012) Projection-like retractions on matrix manifolds. *SIAM J. Optim.*, **22** (1), 135–158.

[45] Lange, K. (2010) *Statistics and Computing: Numerical Analysis for Statisticians*, 2nd edn, Springer, New York.

[46] Yu, D., Won, J.-H., Lee, T. *et al.* (2015) High-dimensional fused lasso regression using majorization–minimization and parallel processing. *J. Comput. Graph. Stat.*, **24** (1), 121–153.

[47] Bierstone, E. and Milman, P.D. (1988) Semianalytic and subanalytic sets. *Inst. Hautes Études Sci. Publ. Math.*, **67**, 5–42.

[48] Bochnak, J., Coste, M., and Roy, M.-F. (1998) *Real Algebraic Geometry*, vol. 36, Ergebnisse der Mathematik und ihrer Grenzgebiete (3), Springer-Verlag, Berlin. [Translated from the 1987 French original, Revised by the authors].

[49] Attouch, H. and Bolte, J. (2009) On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. *Math. Program.*, **116** (1-2, Ser. B), 5–16.

[50] Elsner, L., Koltracht, I. and Neumann, M. (1992) Convergence of sequential and asynchronous nonlinear paracontractions. *Numerische Mathematik*, **62** (1), 305–319.

[51] Lange, K. (2013) *Optimization*, 2nd edn, Springer, New York, NY.

[52] Won, J.-H., Xu, J., and Lange, K. (2019) *Projection Onto Minkowski Sums with Application to Constrained Learning*. International Conference on Machine Learning, pages 3642–3651.

[53] Byrne, C.L. (2008) Sequential unconstrained minimization algorithms for constrained optimization. *Inverse Prob.*, **24** (1), 015013.

[54] Byrne, C.L. (2014) Lecture Notes on Iterative Optimization Algorithms. *http://faculty.uml.edu/cbyrne/IOIPNotesOct2014.pdf*.

[55] Bauschke, H.H. and Combettes, P.L. (2011) *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, vol. 408, Springer, New York, NY.

[56] Ryu, E.K. and Boyd, S. (2016) Primer on monotone operator methods. *Appl. Comput. Math.*, **15** (1), 3–43.

[57] She, Y. (2009) Thresholding-based iterative selection procedures for model selection and shrinkage. *Electron. J. Stat.*, **3**, 384–415.

[58] Bayram, I. (2015) On the convergence of the iterative shrinkage/thresholding algorithm with a weakly convex penalty. *IEEE Trans. Signal Process.*, **64** (6), 1597–1608.

[59] Zhang, C.-H. (2010) Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.*, **38** (2), 894–942.

[60] Juditsky, A. and Nemirovski, A. (2011) First order methods for nonsmooth convex large-scale optimization I: general purpose methods. *Optim. Mach. Learn.*, 121–148.

[61] Helmbold, D.P., Schapire, R.E., Singer, Y., and Warmuth, M.K. (1997) A comparison of new and old algorithms for a mixture estimation problem. *Mach. Learn.*, **27** (1), 97–119.

[62] Azoury, K.S. and Warmuth, M.K. (2001) Relative loss bounds for on-line density estimation with the exponential family of distributions. *Mach. Learn.*, **43** (3), 211–246.