Canad. Math. Bull. Vol. **67** (2), 2024, pp. 415–424
http://dx.doi.org/10.4153/S0008439523000851

Bregman distance regularization for nonsmooth and nonconvex optimization

Zeinab Mashreghi and Mostafa Nasri

Abstract. Solving a nonsmooth and nonconvex minimization problem can be approached as finding a zero of a set-valued operator. With this perspective, we propose a novel Majorizer–Minimizer technique to find a local minimizer of a nonsmooth and nonconvex function and establish its convergence. Our approach leverages Bregman distances to generalize the classical quadratic regularization. By doing so, we generate a family of regularized problems that encompasses quadratic regularization as a special case. To further demonstrate the effectiveness of our method, we apply it on a lasso regression model, showcasing its performance.

1 Introduction

Efficient methods for solving an optimization problem

$$\min_{x\in\mathbb{R}^n} F(x), \tag{1.1}$$

where $F:\mathbb{R}^n \to \mathbb{R}$ is a nonsmooth and nonconvex function, plays a vital role in various domains and are of paramount importance. Numerous studies have been conducted to develop methods for solving the aforementioned optimization problem under reasonable assumptions [3]. The focus of these studies mostly lies in developing strategies to find the optimal solution denoted as x^*, which satisfies the first-order optimality condition

$$0 \in \partial f(x^*),$$

where $\partial f(x)$ is the Clarke subdifferential of f at x [1]. Note that solving $0 \in \partial f(x^*)$ is equivalent to finding a zero of the set-valued operator $\partial f(x^*)$. In addition to the methods based on subdifferentials, penalty function methods offer an alternative approach for solving optimization problems. These techniques involve introducing a penalty term into the original objective function and then employing an iterative method where, at each step, a subproblem is solved within a predefined region [6, 14, 16].

Motivated by [14], the main objective of this paper is to propose a novel approach for solving a class of nonsmooth and nonconvex optimization problems, in which we

Received by the editors July 21, 2023; revised September 21, 2023; accepted October 13, 2023.
Published online on Cambridge Core October 26, 2023.
AMS subject classification: **90C26.**
Keywords: Bregman distance, lasso regression model, local minimizer, nonconvex optimization, nonsmooth optimization.

introduce the utilization of Bregman distance as a key component in constructing a regularized method based on the majorization–minimization (MM) approach. Specifically, we focus on solving the problem defined as (1.1) with

$$F(x) = \frac{1}{2}\|y - Ax\|^2 + \lambda\|x\|_1 - \lambda f_\alpha(x), \tag{1.2}$$

in which $y \in \mathbb{R}^n$ is given, $A \in \mathbb{R}^{m\times n}$ is an $m \times n$ matrix, $\lambda > 0$, and f_α is the Moreau envelope given by

$$f_\alpha(x) = \inf_{z\in\mathbb{R}^n} \left\{ f(z) + \frac{\alpha}{2}\|x - z\|^2 \right\}.$$

The author in [14] proposes a regularization of the form

$$f_\alpha^m(x, w) = f_\alpha(x) - \frac{\gamma_m}{2}\|x - w\|^2 \tag{1.3}$$

to ensure that the subproblem generated at each step of her proposed method has a unique solution. Then, the author uses the regularization to define a majorizer function

$$F^M(x, w) = \frac{1}{2}\|y - Ax\|^2 + \lambda\left[\|x\|_1 - f_\alpha^m(x, w)\right]. \tag{1.4}$$

Then, the author presents an MM method that begins with the initialization of a point, denoted by $x^0 \in \mathbb{R}^n$. Subsequent iterations are then defined by

$$x^{k+1} = \operatorname*{argmin}_{x\in B_{\varepsilon/2^k}(x^k)} F^M(x, x^k) \quad \text{for} \quad k = 0, 1, 2, \dots, \tag{1.5}$$

where $B_{\varepsilon/2^k}(x^k)$ represents a ball centered at the point x^k, with a radius equal to $\varepsilon/2^k$. Note that (1.4) can be obtained when $f_\alpha^m(x, w)$ replaces f_α in (1.2).

The main idea behind the MM approach is to transform a challenging optimization problem into a sequence of simpler and well-behaved subproblems. To solve the subproblems (1.5), there exist various methods proposed in the literature, including those described in [7, 15]. These subproblems are indeed well-behaved and easier to solve because their objective functions are strictly convex and their feasible sets are bounded and convex. By solving the generated subproblems iteratively, a sequence of solutions is obtained that progressively approximates a solution to the original problem. Under certain assumptions, this sequence converges to a point that satisfies the optimality conditions of the original problem.

It is worth noting that the MM method has numerous applications across various fields. Among them, we can mention signal and image processing, support vector machines, nonnegative matrix factorization, and DNA sequence analysis. For a more complete list of applications, the readers can consult [9, 17].

In this paper, we generalize the MM method presented in [14], where we substitute a general Bregman distance for the quadratic regularization term given by (1.3), to be defined in Section 2.

The rest of this paper is organized as follows. In Section 2, we present some basic facts that will be used in this paper as well as our generalized MM method. In

https://doi.org/10.4153/S0008439523000851 Published online by Cambridge University Press

Section 3, we establish our convergence analysis. We present some numerical results in Section 4.

2 Preliminaries and basic facts

In this section, we are going to describe Bregman distances and their properties, and the generalized MM method.

2.1 Bregman functions and distances

Definition Let S be an open and convex subset of $\mathbb{R}^n$, and let $\bar{S}$ be its closure. Consider a convex real-valued function ϕ defined on $\bar{S}$, and let $D : \bar{S} \times S \to \mathbb{R}$ be defined as

$$D(x,w) = \phi(x) - \phi(w) - \nabla\phi(w)^T(x-w).$$

We say that ϕ is a Bregman function and D is its distance induced by ϕ (see [2]) if the following conditions hold.

H1: ϕ is continuously differentiable on S.

H2: ϕ is strictly convex and continuous on $\bar{S}$.

H3: For every $\theta \in \mathbb{R}$, the partial level sets $\Gamma_1(w,\theta) = \{x \in \bar{S} : D(x,w) \le \theta\}$ and $\Gamma_2(x,\theta) = \{w \in S : D(x,w) \le \theta\}$ are bounded for all $w \in S$ and $x \in \bar{S}$, respectively.

H4: If $\{w^k\}_{k=0}^{\infty} \subset S$ converges to w^*, then $D(w^*, w^k)$ converges to 0.

H5: If $\{x^k\}_{k=0}^{\infty} \subset \bar{S}$ and $\{w^k\}_{k=0}^{\infty} \subset S$ are sequences such that $\{x^k\}_{k=0}^{\infty}$ is bounded, $\lim_{k\to\infty} w^k = w^*$ and $D(x^k, w^k) = 0$, then $\lim_{k\to\infty} x^k = w^*$.

Note that when $\{x^k\}_{k=0}^{\infty}$ and w^* are in S, $H4 - H5$ automatically hold true due to $H1 - H3$. Moreover, because ϕ is a strictly convex function, we have that $D(x,w) \ge 0$ for all $x \in \bar{S}$ and $w \in S$, and that $D(x,w) = 0$ if and only if $x = w$.

It is remarkable that, in addition to the Bregman distance, there exists another class of distance defined on the positive orthant of $\mathbb{R}^n$,

$$\mathbb{R}^n_{++} = \{(x_1, \cdots, x_n) \in \mathbb{R}^n : x_i > 0 \ (1 \le i \le n)\},$$

which is formally stated next [11].

Definition Let $\phi : \mathbb{R}_{++} \to \mathbb{R}$ be a strictly convex function. A ϕ-divergence is a function, denoted by $d_\phi : \mathbb{R}^n_{++} \times \mathbb{R}^n_{++} \to \mathbb{R}$, that is defined at (x,y) as

$$d_\phi(x,y) = \sum_{i=1}^n y_i \phi\left(\frac{x_i}{y_i}\right).$$

Some important examples that are known in the literature are given below. We encourage interested reads to consult [4, 5, 10, 11] for further elaboration on Bregman distances and ϕ-divergences as well as more examples.

- Let $S = \mathbb{R}^n$ and $\phi(x) = x^T M x$, where M is an $n \times n$ symmetric and positive definite matrix. In this case, define $D(x,w) = (x-w)^T M (x-w) = \|x-w\|_M^2$.

https://doi.org/10.4153/S0008439523000851 Published online by Cambridge University Press

In particular, if M is the identity matrix, then $D(x,w) = \|x-w\|^2$ reduces to the Euclidean distance squared.

- Let $\phi(t) = t\log(t) - t + 1$ and $0\log(0) = 0$. In this case, the Kullback–Leiber divergence is defined as

$$d_\phi(x,y) = \sum_{i=1}^{n} y_i \phi\left(\frac{x_i}{y_i}\right) = \sum_{i=1}^{n}\left[x_i \log\left(\frac{x_i}{y_i}\right) + y_i - x_i\right].$$

It is worth emphasizing that the Bergman distance may not always be a distance in the usual sense of the term. In general, it lacks symmetry and fails to satisfy the triangle inequality as discussed in [4].

Definition Let $w \in \mathbb{R}^n$ be given. The function $g(\cdot,w) : \mathbb{R}^n \to \mathbb{R}$ is called a local majorizer of a function $h : \mathbb{R}^n \to \mathbb{R}$ at w if

$$h(x) \le g(x,w), \;\; \forall x \in \mathbb{R}^n$$

and

$$h(x) = g(x,w) \;\; \Leftrightarrow x = w.$$

We also say that $g(\cdot,w) : \mathbb{R}^n \to \mathbb{R}$ minorizes f at w when $-g(\cdot,w) : \mathbb{R}^n \to \mathbb{R}$ majorizes $-f$ at w.

Geometrically, our objective is to ensure that the functions h and g are tangent to each other at the point w. In addition, both h and g should have directional derivatives at w. Moreover, we desire that for any small $\|d\|$, the directional gradient of g at w in the direction of d is equal to the gradient of h at w in the direction of d. That is, $\nabla g(w;d,w) = \nabla h(w;d)$ for any small $\|d\|$.

In the classical MM method, the majorizer g^{k-1} is defined as $g^{k-1} = g(\cdot, x^{k-1})$, where g is a function that is tangent to the objective function at x^{k-1}. This majorizer is then minimized over a convex set Ω to obtain the next iterate x^k [9]. That is, $x^k = \underset{x\in\Omega}{\operatorname{argmin}}\, g(x, x^{k-1})$, provided that x^k exists. Then g^k is defined as $g^k = g(\cdot, x^k)$. When the sequence of minimizers $\{x^k\}_{k=0}^{\infty}$ exists, we have the descending property

$$h(x^k) \le g^{k-1}(x^k) = g(x^k, x^{k-1}) \le g^{k-1}(x^{k-1}) = g(x^{k-1}, x^{k-1}) = h(x^{k-1}).$$

2.2 The generalized MM method

In our generalized majorizer–minimizer (GMM) method, which incorporates the Bregman distance, we, respectively, introduce the concepts of the Moreau envelope, regularization, and majorizer functions as

$$f_\alpha(x) = \inf_{z\in\mathbb{R}^n}\{f(z) + \alpha D(x,z)\}, \tag{2.1}$$

$$f_\alpha^m(x,w) = f_\alpha(x) - \gamma_m D(x,w), \tag{2.2}$$

$$F^M(x,w) = \frac{1}{2}\|y - Ax\|^2 + \lambda\left[\|x\|_1 - f_\alpha^m(x,w)\right]. \tag{2.3}$$

https://doi.org/10.4153/S0008439523000851 Published online by Cambridge University Press

It is worth noting that due to the nonnegativity property of D, as given in Definition 2.1, the Moreau envelope F^M acts as a majorizer for the function F. With this in mind, we can now formally present our method as outlined below.

```
GMM Method:
Initialize $x^0 \in \mathbb{R}^n$ and set $\gamma_m > \alpha$.
For $k = 0, 1, 2, \ldots,$ compute
    $x^{k+1} = \operatorname*{argmin}_{x \in B_{\varepsilon/2^k}(x^k)} F^M(x, x^k).$
end
```

We emphasize that in our method, we impose the condition $\gamma_m > \alpha$ to ensure that

$$\frac{1}{2}A^T A + \lambda(\gamma_m - \alpha)\nabla^2\phi(x^k) \succ 0$$

holds, meaning that $\frac{1}{2}A^T A + \lambda(\gamma_m - \alpha)\nabla^2\phi(x^k)$ is a positive definite matrix. We will prove in Lemma 3 that this condition guarantees that $F^M(\cdot, x^k)$ has a unique minimizer within $B_{\varepsilon/2^k}(x^k)$.

3 Convergence analysis of GMM method

This section commences by establishing the foundation for our convergence analysis.

Lemma *Let $x \in \mathbb{R}^n$, $A \in \mathbb{R}^{m\times n}$ and $\lambda > 0$. The function $F : \mathbb{R}^n \to \mathbb{R}$ defined in (1.2) with f_α given by (2.1) is convex if the condition*

$$\frac{1}{2}A^T A - \lambda\alpha\nabla^2\phi(x) \succeq 0$$

is satisfied, where $\nabla^2\phi$ is the Hessian of ϕ that defines the Bregman distance. Moreover, F is strictly convex if

$$\frac{1}{2}A^T A - \lambda\alpha\nabla^2\phi(x) \succ 0.$$

Proof We can write

$$\begin{aligned}
F(x) &= \frac{1}{2}\|y - Ax\|^2 + \lambda\|x\|_1 - \lambda f_\alpha(x) \\
&= \frac{1}{2}\|y - Ax\|^2 + \lambda\|x\|_1 - \min_{z\in\mathbb{R}^n}\{\lambda f(z) + \alpha\lambda\, D(x,z)\} \\
&= \max_{z\in\mathbb{R}^n}\left\{\frac{1}{2}\|y - Ax\|^2 + \lambda\|x\|_1 - \lambda f(z) - \alpha\lambda\, D(x,z)\right\} \\
&= \frac{1}{2}x^T A^T Ax - \lambda\,\alpha\,\phi(x) + \lambda\|x\|_1 + \max_{z\in\mathbb{R}^n} g(x,z),
\end{aligned}$$

where $g(x,z)$ is affine in x defined as

$$g(x,z) = \frac{1}{2}y^T y - \lambda f(z) + \alpha\lambda\phi(z) + \alpha\lambda\nabla\phi(z)(x - z) - \frac{1}{2}(y^T Ax + x^T A^T y).$$

https://doi.org/10.4153/S0008439523000851 Published online by Cambridge University Press

Therefore, F is convex if $\frac{1}{2}A^TA - \lambda\alpha\nabla^2\phi(x)$ is positive semidefinite. Moreover, it is obvious that F is strictly convex if $\frac{1}{2}A^TA - \lambda\alpha\nabla^2\phi(x)$ is positive definite. ■

Lemma *The function f_α^m minorizes f_α for any arbitrary w. Moreover, if $\|d\|$ is small, then it holds that $\nabla f_\alpha(w,d) = \nabla f_\alpha^m(w,w,d)$.*

Proof The fact that f_α^m minorizes f_α can be inferred from (2.3). However, to establish the equality in terms of the directional derivative, a further proof is required. We can write

$$\begin{aligned}\nabla f_\alpha^m(w,w,d) &= \liminf_{\theta\to0^+}\frac{1}{\theta}\left[f_\alpha^m(w+\theta d,w) - f_\alpha^m(w,w)\right]\\ &= \liminf_{\theta\to0^+}\frac{1}{\theta}\left[f_\alpha^m(w+\theta d,w) - f_\alpha(w)\right]\\ &= \liminf_{\theta\to0^+}\frac{1}{\theta}\left[f_\alpha(w+\theta d) - \gamma_m D(w+\theta d,w) - f_\alpha(w)\right]\\ &= \liminf_{\theta\to0^+}\frac{1}{\theta}\left[f_\alpha(w+\theta d) - f_\alpha(w)\right] - \liminf_{\theta\to0^+}\gamma_m D(w+\theta d,w).\end{aligned}$$

To finalize the proof, we only need to show that $\liminf_{\theta\to0^+}\gamma_m D(w+\theta d,w) = 0$. This can be established based on Definition 2.1 and the fact that $\|d\|$ is small. ■

The following theorem provides a sufficient condition that ensures the convexity of $F^M(\cdot,w)$ for every w.

Lemma *F^M is a local majorizer of F at w. Moreover, $F^M(\cdot,w)$ is convex if*

$$\frac{1}{2}A^TA + \lambda(\gamma_m - \alpha)\nabla^2\phi \geq 0$$

holds and that $F^M(\cdot,w)$ is strictly convex if

$$\frac{1}{2}A^TA + \lambda(\gamma_m - \alpha)\nabla^2\phi > 0$$

holds.

Proof Using (2.3), it is clear that the function $F^M(\cdot,w)$ is a local majorizer for F. To prove that $F^M(\cdot,w)$ is convex, we expand F^M and obtain

$$\begin{aligned}F^M(x,w) &= \frac{1}{2}||y-Ax||^2 + \lambda||x||_1 - \lambda f_\alpha^m(x,w)\\ &= \frac{1}{2}x^TA^TAx + \lambda(\gamma_m-\alpha)\phi(x) + \lambda\|x\|_1 + \max_{z\in\mathbb{R}^n} g(x,z,w),\end{aligned}$$

where

$$\begin{aligned}g(x,z,w) = \frac{1}{2}y^Ty - \lambda f(z) - \alpha\lambda(\phi(z) + \nabla\phi(z)(x-z))\\ - \gamma_m\lambda(\phi(w) + \nabla\phi(w)(x-w)) - \frac{1}{2}(y^TAx + x^TA^Ty).\end{aligned} \tag{3.1}$$

https://doi.org/10.4153/S0008439523000851 Published online by Cambridge University Press

Since g is affine in x, F^m is convex if $\frac{1}{2}A^TA - \lambda(\gamma_m - \alpha)\nabla^2\phi(x)$ is positive semidefinite. Moreover, F^m is strictly convex if $\frac{1}{2}A^TA - \lambda(\gamma_m - \alpha)\nabla^2\phi(x)$ is positive definite. ∎

We remark that if $\gamma_m > \alpha$ and H2 hold, then $\frac{1}{2}A^TA + \lambda(\gamma_m - \alpha)\nabla^2\phi > 0$ holds, implying that $F^M(\cdot, w)$ is strictly convex.

The next result states that the GMM method is well-defined and converges.

Theorem *The sequence $\{x^k\}_{k=0}^{\infty}$ generated by the GMM method converges.*

Proof Since

$$x^{k+1} = \operatorname*{argmin}_{x \in B_{\varepsilon/2^k}(x^k)} F^M(x, x^k),$$

we must have

$$||x^{k+1} - x^k|| \le \frac{\varepsilon}{2^k} \quad \forall k = 0, 1, 2, \ldots,$$

which implies that the sequence is bounded. Therefore, by the Bolzano–Weierstrass theorem, $\{x_k\}_{k=0}^{\infty}$ has an accumulation point x^*. Consider a subsequence $\{x^{k_n}\}$ of $\{x^k\}_{k=0}^{\infty}$ such that $x^{k_n} \to x^*$. Fix k and let $k_n > k$. We can write

$$||x^k - x^*|| \le ||x^{k_n} - x^*|| + \mathcal{O}\left(\frac{\varepsilon}{2^{k_n}}\right).$$

Therefore, $\{x^k\}_{k=0}^{\infty}$ converges to x^*. ∎

The following lemma presents a sufficient condition that ensures the strong convexity of the majorizer function F^M.

Lemma *F^M is a-strongly convex if it holds that*

$$\frac{1}{2}A^TA + \lambda(\gamma_m - \alpha)\nabla^2\phi \ge aI. \tag{3.2}$$

Proof Likewise Lemma 3, expand $F^M(x, w)$. Then, one can write $F^M(x, w)$ as

$$F^M(x, w) = \frac{1}{2}x^TA^TAx + \lambda\,(\gamma_m - \alpha)\,\phi(x) + \lambda||x||_1 + \max_{z\in\mathbb{R}^n} g(x, z, w),$$

where $g(x, z, w)$ is the affine function in x given by (3.1). Therefore, F is a-strongly convex in x if (3.2) holds. ∎

Next, we recall a technical lemma that plays a crucial role in the convergence result. This lemma is well-established and applicable to a broad range of functions, provided they possess a local minimizer.

Lemma *If f is a-strongly convex in a set C and $\bar{x}$ is a local minimizer of f, then*

$$a||x - \bar{x}||^2 \le f(x) - f(\bar{x}).$$

https://doi.org/10.4153/S0008439523000851 Published online by Cambridge University Press

Proof See Lemma B5 in [13]. ■

We now proceed to establish our main result, which guarantees the convergence of the GMM method.

Theorem *Assume that (3.2) holds and the sequence $\{x^k\}_{k=0}^{\infty}$ converges to $\bar{x}$. Then $\bar{x}$ is a stationary point for F and $\nabla F(\bar{x}, d) \geq 0$ for every small $\|d\|$. In particular, if $\{x^k\}_{k=0}^{\infty}$ is generated by the GMM method and at each step the majorizer function $F^m(\cdot, x^k)$ is a-strongly convex, then $\{x^k\}_{k=0}^{\infty}$ converges to a stationary point of F.*

Proof Fix x. Since f is continuous and $x^k \to \bar{x}$, we can use (2.2) to conclude that

$$\lim_{k\to\infty} f_\alpha^m(x, x^k) = f_\alpha(x) - \gamma_m D(x, \bar{x}).$$

Therefore, it follows from (1.2) and (2.3) that

$$\lim_{k\to\infty} F^M(x, x^k) = F(x) + \lambda\gamma_m D(x, \bar{x}).$$

On the other hand, apply Lemma 3.6 and use the majorization property to obtain

$$a||x - x^{k+1}||^2 \leq F^M(x, x^k) - F^M(x^{k+1}, x^k) \leq F^M(x, x^k) - F(x^{k+1}).$$

Note that the first inequality is a direct consequence of the definition of the GMM method. Recall that x^{k+1} is always a local minimizer of F^M. By taking the limit on both sides of the aforementioned inequalities as $k \to \infty$, we can express them as

$$a||x - \bar{x}||^2 \leq F(x) + \lambda\gamma_m D(x, \bar{x}) - F(\bar{x})$$

or

$$F(x) - F(\bar{x}) \geq a||x - \bar{x}||^2 - \lambda\gamma_m D(x, \bar{x}).$$

For sufficiently small $\|d\|$, we obtain

$$F(\bar{x} + \theta d) - F(\bar{x}) \geq a\theta^2||d||^2 - \lambda\gamma_m D(\bar{x} + \theta d, \bar{x}). \tag{3.3}$$

On the other hand, H3 implies that there exists $\beta > 0$ such that

$$D(\bar{x} + \theta d, \bar{x}) \leq \beta||(\bar{x} + \theta d) - \bar{x}||^2 = \beta\theta^2||d||^2. \tag{3.4}$$

Therefore, inequalities (3.3) and (3.4) yield

$$F(\bar{x} + \theta d) - F(\bar{x}) \geq a\theta^2||d||^2 - \lambda\gamma_m D(\bar{x} + \theta d, \bar{x}) \geq a\theta^2||d||^2 - \lambda\gamma_m\beta\theta^2||d||^2.$$

As a result,

$$F(\bar{x} + \theta d) - F(\bar{x}) \geq (a - \lambda\gamma_m\beta)\theta^2||d||^2.$$

Dividing both sides of the above inequality by θ yields

$$\frac{F(\bar{x} + \theta d) - F(\bar{x})}{\theta} \geq (a - \lambda\gamma_m\beta)\theta||d||^2.$$

https://doi.org/10.4153/S0008439523000851 Published online by Cambridge University Press

Taking the limit as $\theta \to 0$, we find that

$$\nabla F(\bar{x}; d) = \liminf_{\theta \to 0^+} \frac{F(\bar{x} + \theta d) - F(\bar{x})}{\theta} \geq (a - \lambda \gamma_m \beta) \|d\|^2 \liminf_{\theta \to 0^+} \theta = 0.$$

In particular, if $\{x^k\}_{k=0}^{\infty}$ is generated by the GMM method, then $\bar{x}$ represents a stationary point of F. Using the descending property, the stationary point is a local minimum. ■

4 Application

In this section, we assess the effectiveness of our proposed method by applying the GMM method to a lasso regression model [8]. To evaluate the performance of our approach, we utilize the widely recognized `Credit` dataset, available in the `ISLR` package in R [12]. This dataset consists of 400 observations, where we consider $y =$ `Balance` (average credit card balance in \$) as the dependent variable and `Rating` (credit rating), `Income` (income in \$10,000's), and `Limit` (credit limit) as the independent variables.

To estimate the lasso coefficients of the linear model $y = Ax$, we construct a matrix A with dimensions 400×4. The first column of A contains only 1's, whereas the last three columns correspond to the observed variables `Rating`, `Income`, and `Limit`, respectively. Our goal is to estimate the lasso coefficients for the linear model $y = Ax$, where $x^T = \begin{bmatrix} x_1 & x_2 & x_3 & x_4 \end{bmatrix}$. The lasso coefficients, denoted by $(\hat{x}_L)^T = \begin{bmatrix} \hat{x}_1 & \hat{x}_2 & \hat{x}_3 & \hat{x}_4 \end{bmatrix}$, are obtained by solving the minimization problem

$$\min_{x \in \mathbb{R}^4} \left(\|y - Ax\|^2 + \lambda \|x\|_1 \right).$$

To determine the regularization parameter λ, we employ the `glmnet` package in R and utilize cross-validation.

To present the numerical results, we examine the function F^M defined in (1.4). We evaluate F^M using the MM method in [14], which utilizes the regularization form given in (1.3), and the GMM method, which employs the Bregman distance

$$D(x, w) = (x - w)^T M(x - w) = \|x - w\|_M^2,$$

where M is a 4×4 diagonal matrix with diagonal entries of 10, 11, 12, and 13, respectively. In iteration k, the function $F^M(\cdot, x^k)$ is minimized within the ball $B_{\varepsilon/2^k}(x^k)$. Both the MM method and the GMM method are initialized with parameters $\gamma_m = 2$, $\alpha = 0$, an initial ball with radius of $\varepsilon = 10^{100}$, and an initial point x^0 which is obtained using the ordinary least squares method by applying the `lm` function in R. We employ

Table 1: The lasso coefficients for the `Credit` dataset.

Method	$\hat{x}_1$	$\hat{x}_2$	$\hat{x}_3$	$\hat{x}_4$
`lm`	−342.197	−7.563	0.264	−0.802
`glmnet`	−340.727	−7.446	0.261	−0.769
MM method	−342.621	−7.566	0.263	−0.798
GMM method	−342.370	−7.569	0.264	−0.801

https://doi.org/10.4153/S0008439523000851 Published online by Cambridge University Press

a stopping criterion tolerance of 10^{-4} to determine convergence. The least square coefficients obtained from the `lm` function, the lasso coefficients obtained from the `glmnet` function, as well as the MM method and the GMM method are presented in Table 1. Remarkably, our proposed method yields estimators that closely align with the usual regression coefficients obtained from the *R* software. Moreover, our numerical findings indicate that our proposed method achieves convergence significantly faster than the MM method.

Acknowledgment The work of Zeinab Mashreghi is funded by a grant from the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

[1] J. Bolte, A. Daniilidis, A. Lewis, and M. Shiota, *Clarke subgradients of stratifiable functions*. SIAM J. Optim. **18**(2007), 556–572.

[2] L. Bregman, *The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming*. USSR Comput. Math. Math. Phys. 7(1967), 200–217.

[3] J. Burke, A. Lewis, and M. Overton, *A robust gradient sampling algorithm for nonsmooth, nonconvex optimization*. SIAM J. Optim. **15**(2005), 751–779.

[4] D. Butnariu and A. N. Iusem, *Totally convex functions for fixed points computation and infinite dimensional optimization*, Kluwer, Dordrecht, 2000.

[5] Y. Censor and S. Zenios, *Proximal minimization with d-functions*. J. Optim. Theory Appl. **73**(1992), 451–464.

[6] X. Chen and W. Zhou, *Smoothing nonlinear conjugate gradient method for image restoration using nonsmooth nonconvex minimization*. SIAM J. Imaging Sci. 3(2010), 765–790.

[7] C. Dang and G. Lan, *Stochastic block mirror descent methods for nonsmooth and stochastic optimization*. SIAM J. Optim. **25**(2015), 856–881.

[8] T. Hastie, R. Tibshirani, and M. Wainwright, Statistical learning with sparsity: the lasso and generalizations, Chapman & Hall/CRC Press, New York, 2015.

[9] D. Hunter and K. Lange, *A tutorial on MM algorithms*. Am. Stat. **58**(2004), 30–37.

[10] A. Iusem, *Augmented Lagrangian methods and proximal point methods for convex optimization*. Investigación Oper. **8**(1999), 11–49.

[11] A. Iusem, B. Svaiter, and M. Teboulle, *Entropy-like proximal methods in convex programming*. Math. Oper. Res. **19**(1994), 790–814.

[12] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning with applications in R*, Springer, New York, 2017.

[13] J. Mairal, *Optimization with first-order surrogate functions*. In: ICML 2013 – international conference on machine learning. Vol. **28**, PMLR, 2013, pp. 783–791.

[14] A. Mayeli, *Non-convex optimization via strongly convex majoirziation–minimization*. Canad. Math. Bull. **63**(2020), 726–737.

[15] A. Ruszczyński, *Convergence of a stochastic subgradient method with averaging for nonsmooth nonconvex constrained optimization*. Optim. Lett. **14**(2020), 1615–1625.

[16] I. Selesnick, *Sparse regularization via convex analysis*. IEEE Trans. Signal Process. **65**(2017), 4481–4494.

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

Department of Mathematics and Statistics, University of Winnipeg, Winnipeg, MB R3B 2E9, Canada
e-mail: z.mashreghi@uwinnipeg.ca m.nasri@uwinnipeg.ca

https://doi.org/10.4153/S0008439523000851 Published online by Cambridge University Press